UNIVERSITY OF ALBERTA

COMPARISON OF NUMBER RIGHT, ITEM RESPONSE, AND FINITE STATE APPROACHES TO SCORING MULTIPLE-CHOICE ITEMS

BY

JOYCE L. NDALICHAKO ©

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfilment
of the requirement for the degree of DOCTOR OF PHILOSOPHY

DEPARTMENT OF EDUCATIONAL PSYCHOLOGY

Fall 1997

Edmonton, Alberta

National Library
of Canada

Acquisitions and
Bibliographic Services

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque nationale
du Canada

Acquisitions et
services bibliographiques

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence

Our file Notre référence

The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

L'auteur conserve la propriété du droit d'auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

0-612-23044-9

Canada

UNIVERSITY OF ALBERTA

LIBRARY RELEASE FORM

NAME OF AUTHOR:	JOYCE LAZARO NDALICHAKO
TITLE OF THESIS:	COMPARISON OF NUMBER RIGHT, ITEM RESPONSE, AND FINITE STATE APPROACHES TO SCORING MULTIPLE-CHOICE ITEMS
DEGREE:	DOCTOR OF PHILOSOPHY
YEAR THIS DEGREE GRANTED:	1997

Permission is hereby granted to the University of Alberta Library to reproduce single copies of this thesis and to lend or sell such copies for private, scholarly or scientific research purposes only.

The author reserves all other publication and other rights in association with the copyright in the thesis, and except as hereinbefore provided neither the thesis nor any substantial portion thereof may be printed or otherwise reproduced in any material form without the author's prior written permission.

Ndalichako

University of Dar es Salaam
P.O.Box 35048
Dar-es-Salaam
TANZANIA.

September 5, 1997.

UNIVERSITY OF ALBERTA

FACULTY OF GRADUATE STUDIES AND RESEARCH

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research for acceptance, a thesis entitled COMPARISON OF NUMBER RIGHT, ITEM RESPONSE, AND FINITE STATE APPROACHES TO SCORING MULTIPLE-CHOICE ITEMS submitted by JOYCE LAZARO NDALICHAKO in partial fulfilment of the requirements for the degree of DOCTOR OF PHILOSOPHY.

Rogers

Dr. W. T. Rogers

T Maguire

Dr. T. O. Maguire

Romaniuk

Dr. E. W. Romaniuk

C. A. Norman

Dr. C. A. Norman

A. Brian Nielsen

Dr. A. B. Nielsen

R Traub / T Maguire

Dr. R. E. Traub

Date Sept. 2/97

DEDICATIONS

This work is dedicated to the loving memories of:
my father, Lazaro Elias Ndalichako
and my sister, Deodatha Ndalichako
who passed away during the period of my studies

and

to my mother
Mary Ndalichako
for always loving and believing in me.

ABSTRACT

The oldest form of scoring multiple-choice test items involves awarding one point if the correct option is selected by an examinee and no points if any of the remaining options are selected. However, this number right scoring procedure is criticized (e.g., Shepard, 1993) for its failure to capture different ways in which examinees select the correct. Recently, García-Peréz (1987) proposed the finite state score theory as a method of addressing this issue. An underlying assumption of this approach is the independence of options within an item. The purpose of this study was to compare the finite state score ability estimates with ability estimates yielded by the number right and item response scoring approaches when the option independence assumption is violated.

Responses of 1,232 high school seniors to multiple-choice items contained in an end of year, provincially administered English 30 Examination and a Test of Test-Wiseness were analyzed. The correlations between the pairs of ability estimates in which the finite state scoring approach was used were lower for the subtest of 20 English items which called for the best answer than for the subtest of 48 English items which satisfied the option independence assumption. In agreement, the means and standard deviations for the pairs of absolute differences which involved the finite state ability estimates were greater for the subtest of 20 items than for the subtest of 48 items. The remaining correlations, means, and standard deviations were quite stable across the two English subtests when ability estimates were determined using the number right, one-, two-, and three-parameter item response scoring approaches.

Similarly, the correlations between the pairs of ability estimates which involved finite state ability estimate were lower for the subtest of 18 Test of Test-wiseness items

which contained a pair of opposite or a pair of similar options than for the subtest of 32 items which satisfied the option independence assumption. Further, the means and standard deviations for the pairs of the absolute differences in which the finite state ability estimates were a member were greater for the subtest of 18 items than for the subtest of 32 items. With the exception of the three-parameter ability estimates, the remaining correlations, means, and standard deviations remained stable across the two subtests. A problem that arose was the assignment of zero scores by the finite state scoring algorithm to examinees who scored at or below the chance level.

Item analyses conducted prior to the main analyses did not shed light that could explain the observed differences in correlations, means, and standard deviations of the absolute differences among the ability estimates. It was concluded that lack of option independence and the setting of chance scores to zero adversely influenced the finite state ability estimates. Consequently, the use of finite state scoring approach in scoring multiple-choice items is unjustified. It is recommended that the number right scoring algorithm be used in classroom testing and both the number right and item response scoring approaches be used in large scale testing programs where the sample sizes are adequate and the testing needs are complex.

ACKNOWLEDGEMENTS

I would like to thank many people who have contributed to the completion of this thesis. I would like to express my sincere gratitude to the International Development Research Centre (IDRC) who sponsored my studies. I am also thankful to the University of Dar es Salaam for granting me a study leave. My special thanks to Dr. Naomi B. Katunzi of the University of Dar es Salaam who supported me in many ways before I got the IDRC scholarship. I am also grateful to the Alberta Education, Student and Evaluation Branch, for providing data set used for this study.

I would like to recognize the strong support, encouragement and thoughtful advice I received from Dr. W.T. Rogers, my academic and thesis advisor. He motivated me to work hard even when it was least conducive to do so. Dr. Rogers has been a good model for the pursuit of academic excellence. I admire his sincere dedication and commitment to academic activities.

I would also like to express my appreciations to Drs. T. O. Maguire, L. Stewin, C. A. Norman, A. B. Nielsen, J. P. Das, and E. W. Romaniuk, for their willingness to serve on my thesis committee. I am very thankful to Dr. R. E. Traub of the University of Toronto, who served as an external examiner of my thesis. The comments and suggestions made by all members of my thesis committee contributed greatly to the final product of this work. I appreciate the assistance provided by Dr. Terry Taerum, of the University of Alberta, in solving problems related to the use of SPSS computer program.

The support, encouragement, and understanding of my friends, both near and far, are deeply appreciated. My sincere thanks to Eunice Kanyi who has always been there for

me from the beginning to the end of this journey. My family has greatly contributed to providing a good foundation for my life. My mother, sisters, brothers and relatives have always loved and believed in me and that has given me the courage and confidence to pursue new challenges. Lastly, my heartfelt thanks to my son, Dajos, who always cheered me with a lot of questions: "How was your day, how many more pages do you have to write?" My work is completed and I am looking forward to reading yours.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION 1
  Background Information 1
  Research Questions 6
  Organization of the Thesis 7

CHAPTER 2: LITERATURE REVIEW 9
  Classical Test Theory 10
    Reliability 12
  Validity 13
  Factors Affecting the Reliability and Validity of Test Scores 15
    Effect of Guessing 16
    Credit for Partial Knowledge 19
  Item Response Theory 23
    Assumptions of Item Response Theory 24
    Two-parameter Logistic Model 27
    One-parameter Logistic Model 29
    Three-parameter Logistic Model 30
    Parameter Estimation 32
    Partial Credit Model 34
  Finite State Score Theory 36
    Assumptions of Finite State Score Theory 38
    States of Knowledge 39
    Ability Estimation Methods 43
    Studies Involving Finite State Score Theory 44
  Concluding Remarks 46

CHAPTER 3: METHOD 48
  Data Sources 48
    English 30 Diploma Examination: Reading Section. 49
    Test of Test-wiseness 51
  Data Analyses 53
    Conventional Item Analysis 53
    Item Response Item Analysis 55
      Examination of the Fit of Item Response Models 60
        One-, two,- and three-parameter models 60
        Partial credit model 62
  Ability Estimation 63
    Computer programs 63
  Comparison of Misfitting Items 64
  Agreement Among Ability Estimates 64

CHAPTER 4: RESULTS BASED ON THE ENGLISH 30 EXAMINATION 66
  English 30 Items Which Satisfied the Option Independence Assumption 67
    Subtest Characteristics 67
    Item Characteristics 68
    Conventional Item Analysis Results 71
    Item Response Item Analyses Results 72
      Assessing the Assumptions of Item Response Theory 72
      Examination of the Fit of Items to Item Response Models 76
    Comparison of Misfitting Items Identified by Conventional
      and Item Response Item Response Item Analyses 77
    Comparison of the Ability Estimates 79
  Best Answer English 30 Items 81
    Confirmation of Option Order 82
    Subtest Characteristics 85

Item Characteristics 86
Conventional Item Analysis Results 88
Item Response Item Analysis Results 88
Assessing the Assumptions of Item Response Theory 88
Examination of the Fit of Items to Item Response Models 91
Partial Credit Item Analysis 92
Examination of the Fit of Items to the Partial Credit Model 92
Comparison of Misfitting Items Identified by the Conventional, Item Response, and Partial Credit Item Analyses 93
Comparison of the Ability Estimates 94
Discussion 97

CHAPTER 5: RESULTS BASED ON THE TEST OF TEST-WISENESS ITEMS . . 100
Test-Wiseness Items Which Satisfied the Assumption of Option Independence 100
Subtest Characteristics 100
Item Characteristics 102
Conventional Item Analysis Results 102
Item Response Item AnalysIs Results 105
Assessment of the Assumptions of Item Response Theory 105
Examination of the Fit of Items to the Item Response Models . 108
Comparison of Misfitting Items Identified by the Conventional and Item Response Item Analyses 109
Comparison of the Ability Estimates 110
Test-Wiseness Items With a Pair of Opposite or Similar Options 117
Subtest Characteristics 117
Item Characteristics 118
Conventional Item Analysis Results 119
Item Response Item Analysis Results 120
Assessing the Assumption of Item Response Theory 120

Examination of Fit of Items to the Item Response Models 123
Comparison of Misfitting Items Identified by Conventional and Item Response Item Analyses 123
Comparison of the Ability Estimates 124
Discussion 128

CHAPTER 6: SUMMARY, CONCLUSIONS AND IMPLICATIONS 131
Summary of the Study 131
Background to the Study 131
Purpose of the Study 132
Method of Study 133
Data Analyzed 133
Analyses Conducted 134
Results and Discussion 135
Assumptions of Item Response Theory 135
Examination of the Fit of Items 136
English subtests 136
Test of Test-Wiseness 136
Comparisons of the Ability Estimates 138
English subtests 138
Test of Test-Wiseness 140
Limitations of the study 141
Conclusions and Implications for Practice 142
Implications for Future Research 146

References 148
Appendix A: A Program Used to Compute the Finite State Score Ability Estimates . . 158
Appendix B: Conventional Item Analysis: ENGI Subtest 160
Appendix C: Ratings of the 20 Best Answer English 30 Items: Five Judge 167
Appendix D: Conventional Item Analysis: ENGBD Subtest 169

Appendix E: Conventional Item Analysis: TTWI Subtest 173

Appendix F:Conventional and Item Response Item Analyses for TTWI Subtest:

Removal of 71 Examinees with Chance Scores 178

Appendix G: Conventional Item Analysis: TTWDOS Subtest 181

Appendix H: Conventional and Item Response Item Analyses for TTWDOS Subtest:

Removal of 71 Examinees with Chance scores 184

LIST OF TABLES

Table

1 Table fo Specifications for the English 30 Diploma
Examination, 1992 50
2 Classification of the Test of Test-Wiseness 52
3 Conventional and Item Response Item Analyses Results:
ENGI Subtest 69
4 The First 5 Components: ENGI Subetst 73
5 Comparison of Misfitting Items for Conventional and Item Response Item
Analyses: ENGI Subtest 78
6 Correlations and Mean Absolute Differences: ENGI Subtest 79
7 Judges Ratings Versus Empirical Weighting: ENGDB Subtest 83
8 Conventional, Item Response and Partial Credit Item Analyses Results:
ENGDB Subtest 87
9 The First 5 Components: ENGDB Subtest 89
10 Comparison of Misfitting Items for the Conventional, Item Response,
and Partial Credit Item Analyses: ENGDB Subtest 93
11 Correlations and Mean Absolute Differences Among Ability Estimates:
ENGDB Subtest 95
12 Conventional and Item Response Item Analysis Results:
TTWI Subtest 103
13 The First 5 Components: TTWI Subtest 105
14 Comparison of Misfitting Items: TTWI Subtest 110
15 Correlations and Mean Absolute Differences: TTWI Subtest 111
16 Correlations and Mean Absolute Differences for TTWI: Elimination of
Examinees With Chance Score 114
17 Conventional and Item Analysis Item Analysis Results:
TTWDOS Subtest 119

18 The First 5 Components: TTWDOS Subtest 121

19 Correlations and Mean Absolute Differences Among Ability Estimates:
TTWDOS Subtest 124

20 Correlations and Mean Absolute Differences for TTWDOS:
Elimination of Chance Scores 126

LIST OF FIGURES

Figure 1: Two-parameter Item Characteristic Curves for Four Items 28
Figure 2: One-parameter Model Item Characteristic Curves for Four Items 30
Figure 3: Three-parameter Item Characteristic Curves for Six Items 32
Figure 4: Tree diagram representing possible sequences of events on a 3-option item . . 40
Figure 5: Distribution of Number Right Scores: ENGI Subtest 68
Figure 6: Plot of Eigenvalues Against Component Number: ENGI Subtest 74
Figure 7: The Distribution of Number Right Scores: ENGDB Subtest 85
Figure 8: Plot of Eigenvalues Against Component Number: ENGDB Subtest 90
Figure 9: Distribution of Number Right Scores: TTWI Subtest 101
Figure 10: Plot of Eigenvalues Against Component Number: TTWI Subtest 106
Figure 11: Distribution of Number Right Scores: TTWDOS Subtest 118
Figure 12: Eigenvalues Against Component Number: TTWDOS Subtest 121

CHAPTER 1: INTRODUCTION

Background Information

Despite the current popularity of performance assessments, multiple-choice test items remain a major form of assessment. Multiple-choice test items are commonly used to measure an individual's achievement of knowledge or skills. The use of multiple-choice format in achievement tests and other standardized tests of cognition is based on an understanding that this format provides several advantages over other forms of assessment. These advantages include ease of achieving adequate content coverage, objectivity of scoring, amenability to item analysis, and reliability (Aiken, 1987; Bennet & Ward, 1993; Haladyana, 1994).

In spite of being convenient to use and having some desirable psychometric properties, multiple-choice test items have been criticized for their inability to assess cognition beyond the level of rote memorization. Yamagishi (1991), for example, refers to multiple-choice test items as "a cause of the so-called right-or-wrong way of thinking" (p. 182). Earlier, Choppin (1983) raised the following criticisms against conventional multiple-choice test items:

(a) present the examinee with three or four times as many incorrect statements as correct ones and provide no feedback to help the examinee learn the correct answers;

(b) encourage random guessing; and

> (c) are insufficient in that little information is gained about examinees from their response to a single item. (p. 2)

Bennet and Ward (1993), on another dimension, argued that multiple-choice test items encourage the teaching and learning of isolated facts and rote procedures at the expense of conceptual understanding and the development of problem solving skills. Similarly, Shepard (1993) contended that multiple-choice testing lends itself to multiple-choice teaching. However, such claims have not been universally accepted. Others (e.g., Haladyana, 1997; Hopkins, Stanley, & Hopkins, 1990; Maguire, Hattie, & Haig, 1994; Rothman, 1995) have indicated that, with effort, multiple-choice test items may also measure higher levels of thinking.

Although research shows that multiple-choice test items can elicit complex thinking processes, the way of evaluating thinking processes is limited by the way these items are scored. The oldest form of scoring multiple-choice test items involves awarding one point for the correct response and zero for any other response. However, this number right scoring leads to difficulties when interpreting test scores because a response to a multiple-choice item may come from different "knowledge" levels. These levels include total knowledge, partial knowledge, lack of knowledge, misinformation, and guessing. Glaser (1981) suggested that:

> the task of testing goes deeper than identifying incorrect answers and pointing these out to the students; it should identify the nature of the concept or rule that the student is employing that governs his or her performance in some systematic

> way; in most cases, the student's behaviour is not random or careless, but is driven by some underlying misconceptions or partial knowledge. (p. 926)

In agreement, Rothman (1995) contended that "because students need only choose an answer, not construct it, we do not know whether they possess the knowledge and skills the question was designed to tap or simply guessed well" (p. 54).

Number right scoring has been criticized because it does not capture the full information available in the responses concerning a person's ability (Bock, 1972; Claudy, 1978; Haladyana, 1994; Hambleton, Roberts, & Traub, 1970; Thissen, Steinberg, & Fitzpatrick, 1989; Wainer, 1989). Information about an examinee's level of ability that could be extracted by taking into account which particular incorrect responses are selected is lost when number right scoring is used.

In their critical review of Messick's notion of construct validity, Maguire et al. (1994) questioned the widespread use of scoring models in which a total score is simply the sum of the item scores, with no regard on how examinees arrive at their different total test scores. They viewed tests and scoring procedures as "*mediating empirical operations* to make constructs into numbers, and in that process, it is not obvious that the common practice of administering large numbers of items and computing number correct, or weighted aggregates, retains the representative link between numbers and constructs" (p. 123).

The notion of partial knowledge is based on the assumption that examinees may know enough about the subject matter of an item to eliminate one or more of the options with some certainty. Research has shown that examinees are able to make use of their partial knowledge or their knowledge of test-wise elements when responding to multiple-choice test items (Hutchinson, 1985; Millman, 1966; Rogers & Bateson, 1991a).

Attempts to take partial knowledge into consideration when scoring multiple-choice test items has a long history. Hirtz and Jacobs (cited in Hutchinson, 1985) noted that:

> The assessment of partial knowledge and the control of guessing behaviour have been two goals of measurement specialists since the introduction of the objective test format. ... Related to the problem of tendencies to omit or guess at items and the problem of chance success under answer-every-item instruction is the evaluation of correct or incorrect response to an objective item. The extent of a respondent's knowledge concerning item *i* cannot be accurately assessed using conventional scoring procedures; an incorrect response is not sufficient evidence to conclude nothing is known concerning the item. The correct response is also insufficient evidence for the converse. (p. 4)

Clearly, there is a need for improved methods of scoring multiple-choice items. Toward this end, a number of scoring methods have been developed. These methods include formula scoring, answer-until-correct, confidence scoring, and option weighting scoring (Smith, 1987). All methods are based on the concept of an item response that is broader than simply scoring correct options. Incorrect responses, as well as correct responses to multiple-choice items are used to extract more information about examinees' states of knowledge.

Following his review of the answer-until-correct, option weighting, multiple correct options, confidence testing, and subset selection methods of scoring multiple-choice items, Frary (1989) concluded that these methods failed to extract all the potentially available information from multiple-choice responses. He noted that the main problem with these methods was centred on the assumptions made about examinees' behaviour when responding to a multiple-choice item. García-Peréz and Frary (1991b) summarized the situation succinctly: "in most cases the assumptions made are local to the response mode that the model was designed for, thereby failing to provide a comprehensive picture of test response behaviour" (p. 274).

An alternative approach involves the use of item response theory (Lord, 1952, Lord & Novick, 1968). Proposed initially by Lord (1952), the initial item response scoring models involved only the correct option. Subsequently, the scoring models were extended to incorporate incorrect options. These latter models include the partial credit model (Masters, 1982), the nominal response model (Bock, 1972), and the graded response model (Samejima, 1969). These models require item options to be ordered in terms of degree of correctness. Since most test items do not meet this requirement, these models are not widely used. Furthermore, the complexity of such models limits their applications to large scale testing programs.

Scoring methods which are capable of recovering additional information from examinees are still being sought. Recently, García-Peréz (1987) proposed a method based on the use of the finite state score theory (Hutchinson, 1982) in an attempt to address the limitations of previously developed scoring methods. With this approach, information from each option is used. Zin (1991), in her review of methods devised in an attempt to account for partial knowledge and guessing, noted that the finite state score theory seems to be a promising trend in the area of measurement, and called for continued testing of the theory to gain a greater understanding of the utility of finite state scoring procedures (p. 10).

Research Questions

This study was designed to answer the following research question:

To what extent are ability estimates from finite state scores, number right, and item response models similar to each other when the assumption of independence of options made in finite state score theory is violated?

The derivation of the finite state scoring polynomials used in the finite state score estimation procedure is based on the assumption that each option is classified independently by an examinee. Consequently, a test to be scored using finite state scoring polynomials should not have two options whereby one is simply the negation or the opposite of the other or two options which are similar to each other. Furthermore, items eliciting the best answer have to be avoided when this theory is used because one option cannot be classified without having read the others. Many tests contain such items and options (Rogers & Bateson, 1991, Sarnacki, 1979). Therefore, it is essential to study the behaviour of the finite state score theory when the assumption of independence is violated in order to assess its utility in the context where such items are commonly used.

Ancillary research question. Interpretation of the differences, if any, among finite state scores, number right scores, and item response ability estimates may be facilitated by the knowledge of item analysis statistics for the items from which scores are derived. Consequently, item analyses were completed using conventional item analysis procedures and those based on item response theory. Both approaches were considered in order to examine whether they will lead to the same conclusion regarding item quality. Previous research in which several approaches have been used in item analysis show that the results are not always the same (Knodel, 1974). Thus, an ancillary research question addressed in the present study was:

> *To what extent do conventional and the item response item analyses lead to the same decisions regarding item quality?*

Organization of the Thesis

Issues related to the use and scoring of multiple-choice test items have been introduced and discussed in Chapter 1. Specific research questions addressed are presented at the end of the chapter together with the rationale for each question. In Chapter 2, the classical test theory, item response theory, and the finite state score theory are examined in terms of their main concepts and underlying assumptions. The Chapter is concluded with a review of studies involving methods of scoring multiple-choice items. Methods of investigation used in this study are presented in Chapter 3. Analytical methods and computer programs used to compute various ability estimates are provided for each of the scoring models.

The results from the analyses described in Chapter 3 are presented in Chapters 4 and 5. Presentation of results is organized in terms of the four subtests analyzed. In Chapter 4, the results of the first two subtests consisted of items selected from an end-of-year English 30 Examination administered provincially to high school seniors are presented. The first subtest contained 48 items which satisfied the assumption of option independence; the second subtest consisted of 20 items which, being the best answer rather than correct answer items, failed to satisfy the assumption of option independence.

In Chapter 5, the results of two subtests formed from the Test of Test-Wiseness (Rogers & Wilson, 1993) are presented. The first subtest contained 32 items which satisfied the assumption of option independence, while the second contained 18 items which did not satisfied this assumption because of the presence of similar options or opposite options (Millman, 1966). In each of these chapters, the results for each subtest are presented in terms of the psychometric characteristic of the items and the fit of items to the models. Finally, the results of the comparison of models made at the subtest level in terms of the correlations among the pairs of estimates and the agreement among the solutions examined using the mean absolute differences among the pairs of transformed ability estimates are presented.

In Chapter 6, summary of the study, limitations of the study, conclusions and implications for practice are presented. The implications for future research are also presented.

CHAPTER 2: LITERATURE REVIEW

The literature review is organized in three main sections. In the first section, a brief overview of classical test theory, in which the conventional item analysis is rooted, is presented. This is then followed by a discussion of formula scoring and scoring methods that take into account partial knowledge. Item response theory is presented in the second section, followed by the presentation of the finite state score theory in the third section.

Before presenting the classical test theory, it is pertinent to mention that there are two main frames-of-reference within which test scores are interpreted. These are the norm-referenced frame-of-reference and the criterion-referenced frame-of-reference. Norm-referenced interpretations are focussed on the level at which an examinee falls on a specific trait. Criterion-referenced interpretations are focussed on whether an individual or a group has mastered a specific set of educational objectives.

Norm-referenced testing makes possible the comparison of an examinee's test performance to that of the norm group to which he or she belongs. Much like a survey instrument, norm-referenced tests tend to provide a very limited view of what students know and can do. Generally, representing performance in comparison to that of other students says little about the skills and knowledge a student has attained (Rothman, 1995, p. 53-54). Conversely, criterion-referenced testing, with its emphasis upon a specified objective, shows what a person knows or can do independent of the way anyone else performs on the test (Traub, 1994, p. 140).

For the purposes of this study, the term test score is used in the context of norm-referenced test scoring. Consequently, the literature reviewed in this chapter is confined to norm referenced testing.

Classical Test Theory

Classical test theory (Gulliksen, 1950) was an early attempt at developing a mathematical approach to mental measurement and it has maintained a strong influence among testing and measurement practitioners. The theory attempts to establish the relationship between the observed score and the true score.

The observed score for an individual represents the ability of that particular individual on a particular sample of items (Suen, 1990). The true score for an individual *j* is defined as *j's* mean observed score on an infinite number of parallel or interchangeable tests. An error score is defined as the difference between the observed and true scores. It can either be positive or negative. If an error score is positive then an individual's true score will be overestimated by the observed score. Conversely, a negative error score means that an individual's true score is underestimated by the observed score.

Clearly, the definition of true score indicates that the true score is a theoretical concept since the construction and administration of parallel forms of a test to the same subject is not feasible. Thus, to obtain estimates of the true score and the error of measurement, measurement specialists (e.g., Spearman, 1904; Lord & Novick, 1968) turned to the variation among individuals within the population of interest. Five assumptions were required to do so (see Allen & Yen, 1979, pp. 57-59). The first two assumptions correspond to the model for an individual j:

$$X_{jf} = \tau_j + \epsilon_{jf} ,$$

where X_{jf} is the observed score of person j on form f, τ_j is the true score of person j, and ϵ_{jf} is the error score for person j on form f, and

$$\xi(X_{jf}) = \tau_j .$$

The second assumption postulates that for person j, the true score (τ_j) is the mean of the theoretical distribution of the observed scores (X_{jf}) that would be found in repeated independent testing of the same person j with parallel forms of the same test. As previously mentioned, since it is impossible in practice to have infinite number of parallel forms of testing, true score remains a theoretical entity.

The next three assumptions pertain to a population of examinees administered one form of a test. The third assumption states that the error scores and the true scores obtained from a population of examinees on one test are uncorrelated:

$$\rho_{\tau\epsilon} = 0 ,$$

where $\rho_{\tau\epsilon}$ is the correlation coefficient.

The fourth assumption concerns the independence of error scores. It is assumed that the error scores on two parallel forms of a test are uncorrelated:

$$\rho_{\epsilon_1\epsilon_2} = 0 .$$

The fifth assumption is that, as the number of people administered form *f* of the test increases, the mean error score approaches zero:

$$\bar{\epsilon}_f = \lim_{N \to \infty} \sum_{j=1}^{N} \frac{\epsilon_{jf}}{N} = 0 .$$

where N is the number of people, and $\bar{\epsilon}_f$ is the mean error score on form f.

Reliability

Given both true scores, τ_j and error scores, ϵ_{jf} are not observable, the focus of classical test score theory was directed toward the estimation of these parameters. Spearman (1904) noted that given the assumptions made above are satisfied, the variance of the observed scores for a population of examinees, σ_x^2 on form f, is equal to the variance of the true scores, σ_τ^2, plus the variance of the error scores, σ_ϵ^2:

$$\sigma_x^2 = \sigma_\tau^2 + \sigma_\epsilon^2 .$$

For convenience reasons, the designation of the form f has been deleted here and in the remaining part of this thesis (i.e., $\sigma^2_{\epsilon f} = \sigma^2_\epsilon$). Spearman defined the ratio of the true score variance to the observed score variance as the reliability of the test scores, ρ_{XX}:

$$\rho_{XX} = \frac{\sigma_\tau^2}{\sigma_X^2}$$

$$= \frac{\sigma_\tau^2}{\sigma_\tau^2 + \sigma_\epsilon^2} :$$

Several methods - parallel forms, test-retest, and internal consistency - have been developed to estimate reliability. Unfortunately, they differ from one another due to the ways different sources of measurement errors are estimated (see, for example, Stanley, 1971). But, given a value for the reliability, an estimate of the variance error of measurement can be found:

$$\sigma_\epsilon^2 = \sigma_x^2(1 - \rho_{XX}).$$

If the reliability is one, it follows that the error and variance of the error of measurement are equal to zero and the obesrved score equals the true score. However, this "ideal" situation is never met. No matter how well tests and test items are developed, scores obtained from a measuring instrument will contain some error of measurement. However, if reliability is high, then the variance of the error of measurement will be low and individual errors of measurement will be small.

Validity

It is quite possible for a test to yield consistent scores that are lacking validity (Ary, Jacobs, & Razavieh, 1996; Hopkins et al., 1990). Therefore, there is a need to determine what the scores on a test mean and how they are interpreted. The manner in which this need is addressed goes beyond analytic procedures associated with the classical test score theory (i.e., beyond reliance upon empirical correlational and predictive evidence) (Cronbach, 1971; Gulliksen, 1950; Messick, 1989).

The most commonly used definition of validity today is the definition provided by Messick (1989). He defined validity as "an integrated evaluative judgement of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of the inferences and *actions* based on test scores or other modes of assessment" (p. 13, emphasis in the original). He then pointed out that validity evidence is normally gathered through a construct-validity study or a series of construct validation studies in which both logical and empirical approaches are combined (Messick, 1989). Correlational, experimental, and quasi-experimental studies, content analysis, use of expert panels and protocol analysis are the major approaches used to gather the needed evidence. For example, convergent and discriminant validity evidence can be derived from a set of correlations among tests purposely selected: tests measuring similar concepts are expected to correlate highly, thereby providing convergent validity evidence, while tests measuring dissimilar concepts are expected to correlate lowly, thereby providing discriminant validity evidence. Evidence from experiments and quasi-experiments that agrees with hypothesized results also provides convergent validity evidence. Evidence collected from content analysis and analysis of individual "think aloud" protocols can be used to confirm the extent to which a psychological construct is reflected in the responses of individuals.

Ary et al. (1996) and Feldt and Brennan (1989) observed that many measurement specialists accord greater attention to reliability than to validity. One of the possible explanations for this is that it is easier to determine reliability using mathematical formulae. Furthermore, the estimation of reliability is possible using only the test data while the evidence for or against validity requires external data.

Reliability has also received greater attention than validity because it can be measured more objectively. Subjective judgement plays a significant role in determining the validity of a test score. Questions about the adequacy of criteria, the proper definitions of human constructs, the appropriateness of test content, and the implications of the test scores inevitably involve subjective judgement (Feldt & Brennan, 1989). In their concluding comments, Feldt and Brennan (1989) asserted that "a game played by subjective, rather than mathematical rules, may be harder to play well, more prone to professional controversy, and attract fewer players" (p. 143). However, they concluded their chapter by acknowledging the primacy of validity in the evaluation of the adequacy of an educational measure: "no body of reliability data, regardless of the elegance of the methods used to analyze it, is worth very much if the measure to which it applies is irrelevant or redundant" (p. 143).

Factors Affecting the Reliability and Validity of Test Scores

There are many factors that serve to confound reliability and validity. Commonly included among these factors are heterogeneity of variance, time limits, item characteristics, and test length (Crocker & Algina, 1986; Traub, 1994, pp. 98-110). Strictly speaking, these factors affect the reliability of test scores. However, since reliability is a necessary but not sufficient condition for validity, implicitly, these factors affect the validity of test score interpretations as well.

The value of a reliability coefficient depends on the magnitude of the variance of true scores; given constant observed score variance, the higher the true score variance the greater the reliability coefficient. Similarly, the longer the test the larger the reliability coefficient. Item characteristics that have an effect on the value of a reliability coefficient are the item-total correlation or discrimination of the item, and the proportion of examinees who answered the item correctly, the difficulty of the item or p-value. The greater the indices of discrimination for a set of items, the larger the expected value of the reliability coefficient. Item difficulty affects the coefficient of reliability in the sense that too easy or too difficult items do not differentiate between low- and high-achieving examinees. This, in turn, results in lower discrimination indices and, consequently, low reliability coefficients (Sax, 1989).

Some measurement specialists (e.g., Ebel & Frisbie, 1991; Feldt, 1993; Lehmann & Mehrens, 1991; Sax, 1989) have observed that although there might be a small reliability advantage for a test with items whose p-values are concentrated around the optimum difficulty level, there may be a genuine psychological advantage to be gained by including a few easy items at the beginning of a test and a few difficult ones at the end, with the rest being of average difficulty. Examinees of low and average ability may gain confidence when they are able answer the first few questions, despite having to struggle with the few difficult ones at the end of the test.

In addition to the factors discussed in the previous paragraphs, other factors that affect both reliability and validity of test scores are guessing and the impact of partial knowledge. The two factors are of particular concern in the present study.

Effect of Guessing

When individuals are responding to a test item, there is a possibility that if they do not know the answer, they may decide to guess. Although the problem of guessing occurs in all test item forms, its potential effects are more amenable to analysis in selection items (e.g., true-false and multiple-choice items) in which there are a finite number of alternatives from which to make a guess.

Guessing "interferes with what would seem to be a major goal of testing - namely, to extract the test taker's true ability from overt responses to the test" (Budescu & Bar-

Hillel, 1993, p. 279). This interference has led to the development of several different "correction" procedures. The oldest of these is formula scoring (Frary, 1988). Formula scoring method is based on the assumption that if examinees do not know the correct answer to a test item, they will guess randomly among all of the options. The corrected score, given by the formula below, is considered a better estimate of an individual's "true" score than uncorrected scores:

$$X_{jc} = R_j - \frac{W_j}{k-1} ,$$

where X_{jc} is the score for person j corrected for guessing, R_j is the number of correct answers and W_j is the number of incorrect answers initially obtained by person j, and k is the number of alternatives per item. The omitted items are not regarded as wrong items when formula scoring is used[1]. The underlying logic in using formula scoring is to deprive the examinee of the number of points which are estimated to have been gained from random guessing (Crocker & Algina, 1986).

The use of the formula scoring rests on the following two assumptions: (a) all guesses are random, with each response having an equal probability of occurring, and (b) every incorrect response results from guessing (Thorndike, Cunningham, Thorndike, & Hagen, 1991). The first assumption ignores the possibility that examinees sometimes answer questions on the basis of partial knowledge by ruling out one or more of the options as impossible. The second assumption precludes the possibility that an examinee might make an incorrect response due to misinformation.

---

[1] Sometimes a correction is done for the omitted items. The corrected score for the omitted items is given by: $X_c = R + O/k$ where O is the number of omitted items. This correction increases an examinee's observed score by awarding additional points for the omitted items.

Contrary to the first assumption, Angoff and Schrader (1986), for example, noted that examinees who make guesses are operating on the basis of partial knowledge (p. 242). Similarly, Rogers and Bateson (1991a) observed that examinees rarely guess "blindly" and that they make use of partial knowledge and test-wiseness to arrive at the answers they choose. Similar views were extended by Mehrens and Lehmann (1991) in their argument that "to the extent that guessing occurs in multiple-choice test items, logic and evidence suggest that it is informed guessing that predominates" (p. 461).

In their review of studies in which formula scoring results were compared to non-formula scoring results, Diamond and Evans (1973) found that reliability estimates were similar or slightly higher for the corrected scores than in the non-corrected scores. Empirical validity coefficients computed using corrected scores tended to be slightly higher than when uncorrected scores were used, though the differences were small. In agreement, Mehrens and Lehmann (1991) argued that luck in guessing may raise or lower a particular student's score by a mark or two, but it will not make any noticeable difference unless the test is unreasonably short. Their argument implies that guessing should not influence the psychometric quality of a test if the test is long.

Angoff (1989), on the other hand, investigated whether guessing really helps or not. His findings suggested that the advantages of guessing depend, at least in part, on the ability of the examinee. Rogers and Bateson (1991a) supported this finding. In their study they found that while examinees guessed, they did so only after attempting to delete some options. They further reported that examinees with greater ability were more successful in guessing than those with lower ability.

The use of formula scoring may lead to undesirable variance due to personality characteristics. This is well summarized by Rowley and Traub (1977):

> If one encourages students to answer all questions whether they know the answer or not, a source of random variance is introduced which decreases both reliability and validity; on the other hand, if one attempts to discourage students from guessing [by using formula scoring], it is apparent that some students will comply to a greater extent than others, causing the test results to be contaminated by personality factors which the test was not intended to measure. (pp 16-17)

Credit for Partial Knowledge

The concept of awarding partial credit stems from the belief that the distractors are differentially informative. Haladyana (1994) pointed out the importance of taking into consideration this information when scoring multiple-choice test items. He acknowledged that "...even though we do not intend to write items with differential distractor effectiveness, reliable information exists in the distractor responses that can, productively, be used to score tests" (p. 18). Researchers who have used the partial credit model have shown that the degree of correctness of an answer can be quantified and used as an additional source for estimating an individual's ability (Smith, 1987; Bock, 1972).

While the general intent of formula scoring is to prevent examinees from receiving "undeserved" points, the method does not take into account the partial knowledge. Consequently, considerable efforts have been devoted to determining ways to award credit for partial knowledge. These efforts were stimulated by dissatisfaction with the all-or-nothing character of number right and formula scoring of multiple-choice test items. Crocker and Algina (1986) grouped scoring procedures that account for partial knowledge into three main classes: confidence weighting, answer-until-correct, and option weighting.

Confidence testing. Confidence testing (De Finetti, 1965) is a scoring method in which an examinee is asked to select the perceived option most likely to be correct and, at the same time, express the degree of confidence in that choice. Confidence testing is scored in such a way that two examinees choosing the same response may receive different scores for that item depending on their indication of the degree of confidence in their responses.

A review of studies that have employed this method can be found in Frary (1989). The results of the studies show an increase in reliability at the expense of additional time required for administration and scoring. For example, Hakistan and Kansup (cited in Frary, 1989, p. 88) found that confidence testing "resulted in improved reliability, but the additional testing time required for confidence testing would have permitted lengthening the number-right test sufficiently to overcome this difference." Similarly, Pugh and Brunza (1975) found improved reliability but not validity when they applied confidence testing. Frary (1989) concluded his review as follows:

> it appears that confidence testing has found a narrow niche in the world of measurement. Moreover, it seems that its use is more appreciated for secondary

> qualities than for possible improvement of the psychometric characteristics of scores. (p. 89)

Answer-until-correct. The answer-until-correct procedure presents students with instant feedback on their response to an item. If the response of a student is correct, then the student is directed to continue to the next question; if the response is incorrect, then the student is asked to attempt the item again and to choose among the remaining options. This testing procedure must be used with special rub-out type or latent image answer sheets or with computerized testing so that a record is made of the number of responses attempted for each item. Answer-until-correct tests are scored by subtracting the total number of responses made by an examinee from the total number of possible responses (Gilman & Ferry, 1972).

An important advantage of the answer-until-correct method is that the immediate feedback inherent in the process may enhance learning. Moreover, the method provides students with an opportunity to examine their incorrect responses to determine why they are incorrect and reevaluate the remaining options in the light of that information. However, Hanna (1975) found that while the answer-until-correct scoring method lead to increased reliability, it seemed to decrease validity .

Option weighting. The option weighting method is based on the assumption that item response options vary in the degree of correctness, and examinees who select a "more correct" response have more knowledge than those who select "less correct" responses (Crocker & Algina, 1986). Option weighting differs from number right scoring in that each option is weighted according to its degree of correctness. Weights may be obtained through two general types of procedures, *a priori* and *empirical* (Smith, 1987).

The *a priori* procedure is based on judges' ratings of the conceptual degree of correctness for each option. The *empirical* procedure uses data from some standard administration to develop the response weights. Point-biserial correlations or p-values are normally used to rank the options.

An early study in which the empirical method was used was completed by Davis and Fifer (1959). They used point-biserial correlation coefficients to weigh the options. Their results showed a significant gain in reliability with a minor decrease in validity. In agreement, Claudy (1978) weighted options using point-biserial correlations and found the method to yield high estimates of internal-consistency reliability. Similarly, Downey (1979) found that while option weighting resulted in an increase in reliability coefficient, the criterion-related validity coefficient decreased (cited in Crocker & Algina, 1986, p. 407).

The use of partial credit scoring methods reviewed - confidence testing, answer-until-correct and option weighting - have resulted in very small increases in test score reliability and sometimes decreases in test score validity. It appears that these methods provide no consistent potential for worthwhile improvement of the psychometric properties of test scores. Therefore, researchers are still in search for better methods of taking into account the partial knowledge.

The issues of guessing and partial knowledge have also been addressed in item response theory, a measurement framework used extensively in large scale testing programs. For example, a guessing parameter is included in what is known as the three-parameter item response model (Lord, 1980). The main ideas in item response theory are highlighted in the next section, followed by an examination of the model for partial credit scoring based on this theory.

Item Response Theory

Item response theory has gained wide acceptance among educational practitioners. Proponents of this theory argue that it has revolutionalized the field of educational measurement. The theory was developed independently by Lord (1952) and Rasch (1960). Based upon the normal ogive, Lord's work proved intractable because of the complex mathematics involved in obtaining a solution. But Birnbaum (1968) suggested the use of logistic models which made the application of Lord's work possible.

In item response theory, the amount of a trait or ability possessed or achieved by an examinee is related to the probability of answering correctly an item. Item response theory focuses on the responses to individual test items rather than total test scores. Using complicated mathematical models, item response theory provides a mathematical equation that predicts the probability of a correct response to a test item as a function of an examinee's ability and certain characteristics of the item.

Assumptions of Item Response Theory

The power of item response theory resides in the potential of making inference about performance at the item level. Although item response theory offers many advantages over classical test theory, these advantages are gained at the expense of strong assumptions about the nature of data used. If these assumptions are not met, the advantages associated with the use of item response theory may not be fully realized (McKinley & Mills, 1985).

The first assumption concerns the dimensionality of the ability (latent trait) space. It is assumed that the probability of a correct response by an examinee can be attributed to his or her standing on a specific number of latent traits or abilities. In most applications of item response theory it is assumed that the latent space is unidimensional. In this case it is implicit that the examinee performance can be accounted for by a single latent trait.

Most cognitive achievement tests measure, to different degrees, multiple skills (Ackerman, 1989; Hambleton, Swaminathan, & Rogers, 1991; Rackase & McKinley, 1991; Traub, 1983). Thus, measurement specialists have emphasized that real test data, most often, cannot be well modeled by a unidimensional model. As a result, unidimensionality is considered a "fragile assumption" (Traub, 1983, p. 59).

Currently, the conceptualization of unidimensionality has changed. Emphasis is now placed on a dominant dimension rather than one dimension. For example, Hambleton et al. (1991) contended that "the assumption of unidimensionality cannot be strictly met because several cognitive, personality, and test-taking factors always affect test performance" (p. 9). They suggested, however, that provided there is one dominant factor underlying performance, the weaker dimensions can be safely ignored. Similarly, Nandakumar (1994) maintained that in order to essentially satisfy the assumption of unidimensionality, a given test must measure a single dominant factor (p. 17). Otherwise, multidimensional models should be used.

The second assumption is that of local independence. That is, the item responses of a given examinee are statistically independent. This means an examinee's performance on one item does not affect his or her performance on the other items in the test (Hambleton & Cook, 1977, pp. 77-78). When this assumption is satisfied, the probability of any pattern of item scores for an examinee is simply the product of the probabilities of individual item response probabilities. When the assumption of unidimensionality is met, local independence is obtained (Hambleton et al., 1991, p. 11). However, local independence can be met even when the data set is not unidimensional as long as all the ability dimensions influencing performance have been taken into account. For example, local independence may not hold when a test item contains a clue to the correct answer. In this case, some examinees will detect the clue and some will not. The ability to detect the clue is a dimension other than the ability being tested. Thus, if a unidimensional model is fitted, local independence will not be obtained.

The third assumption is that the relationship between a person's ability level and performance on an item can be mathematically described by means of an item characteristic curve. An item characteristic curve is a mathematical function that relates the probability of success on an item to the ability measured by the item set or test that contains it. Essentially, an item characteristic curve represents the examinee's probability of success as a function of ability.

The mathematical form of an item characteristic curve is an important point of distinction among different item response models. Two aspects are considered: the number of parameters, and the mathematical function. In parametric functions of unidimensional item response models the number of item parameters in the model is one, two, or three, and the corresponding models are named according to the number of parameters. The first parameter is the difficulty of an item, the second is the discrimination and the third parameter accounts for guessing. The models include an ability parameter in addition to the item parameters. The ability parameter, commonly referred to as "theta", θ, is a non linear transformation of the number right scores. The ability θ is estimated using the maximum likelihood estimation procedure as described in this chapter under the section "parameter estimation."

There are two popular mathematical functions - the normal ogive (Lord, 1952) and the logistic function (Birnbaum, 1968). The logistic function was introduced to overcome the intractability of obtaining a solution using the normal ogive (Birnbaum, 1968). The logistic function is more convenient to work with because it is an explicit function of item and ability parameters while the normal ogive function involves integration.

The one-, two- and three-parameter models are presented in the next three subsections. The two-parameter model (Lord, 1952; Lord & Novick, 1968) is presented first since the others can be easily derived from it.

Two-parameter Logistic Model

Based on the assumptions that item difficulty and discrimination parameters vary and that there is no guessing, the probability of a correct response in the two-parameter logistic function is given by:

$$P_i(\theta) = \frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}} \quad i = 1, 2, ..., n ,$$

where $P_i(\theta)$ is the probability that a randomly chosen examinee with ability θ answers item i correctly; b_i is the item difficulty index, the point on ability scale where the probability of a correct response is 0.5; a_i is the discrimination index, the value which is proportional to the slope of the item characteristic curve at the point b_i; D is a scaling factor which is commonly set to 1.7 so as to make the logistic function as close as possible to the normal ogive function; and n is the number of items.

Theoretically, difficulty values can range from -∞ to +∞; in practice, values are usually in the range of -4 to +4. Items with high values of *b* are difficult item; low ability examinees have low probabilities of responding correctly to an item with a high value for b. Similarly, the discrimination parameter can range from -∞ to +∞. However, in practice, typical values range from 0 to 2. The higher the *a* value, the more sharply the item discriminates among examinees at different ability levels. Examples of item characteristics curves for the two-parameter model are shown in Figure 1.

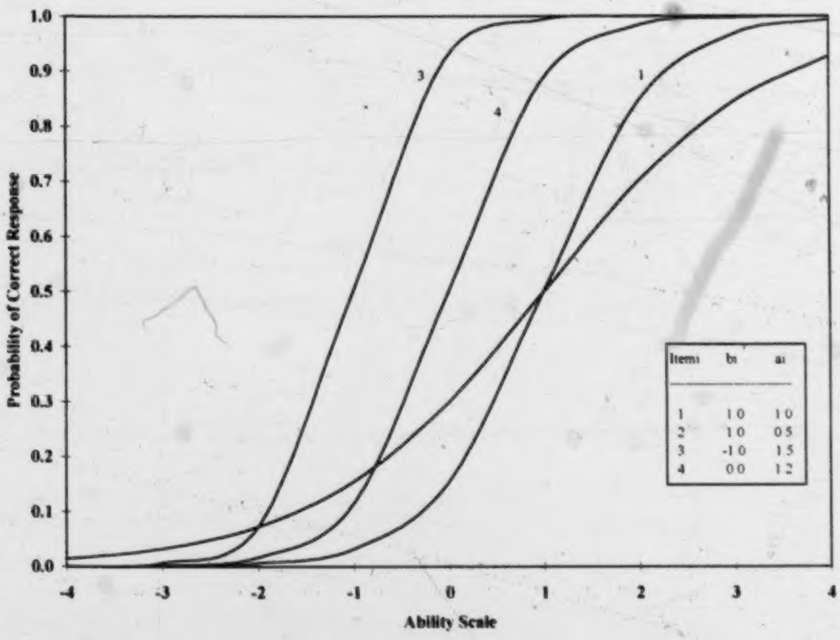


Figure 1: Two-parameter Item Characteristic Curves for Four Items
Redrawn from: Hambleton et al. (1991, p. 16)

As shown in Figure 1, items 1 and 2 are equally difficult; items 3 and 4 differ from each other, with item 3 less difficult than item 4. Item 3 has a higher discrimination index than item 2. That is, item 2 discriminates less well than item 1 as seen from its "flatter" curve. As reflected by the curves for items 1, 3, and 4, these items differ from each other in terms of difficulty parameter. For item 3, which is the easiest item, examinees with at least 0 ability have a probability of .90 of answering correctly this item. For item 1, the most difficult item, examinees with at least 0 ability have a probability of only .15 of answering this item correctly. Each curve has a lower asymptote of zero, reflecting the fact that in the two-parameter model it is assumed that there is no guessing.

Thus, examinees with very low ability have a zero probability of correctly answering the item.

One-parameter Logistic Model

The one-parameter model is based on the additional assumption that items have equal discriminating power. Under this model, the items vary only in terms of difficulty. This model specifies the probability of a correct response to an item as an exponential function of a person's ability, θ, and item difficulty b_i:

$$P_i(\theta) = \frac{e^{Da(\theta - b_i)}}{1 + e^{Da(\theta - b_i)}} \qquad i = 1, 2, ..., n,$$

where $D = 1.7$ and a, the item discrimination, is constant for each item.

Rasch (1960), working independently of Lord (1952), developed the one-parameter model as well. For this reason, the one-parameter model is also known as the Rasch model. Using Rasch notations, the formula for the one-parameter model becomes:

$$P_i(\theta) = \frac{e^{(\theta^* - b^*_i)}}{1 + e^{(\theta^* - b^*_i)}} \qquad i = 1, 2, ..., n ,$$

where $\theta^* = Da\theta$ and $b_i^* = Dab_i$ (Rasch, 1960). Written in this form, the formula more clearly shows that with the one-parameter model, the proportion of examinees responding correctly to an item is a function of the examinees' ability and the difficulty of the item. Examples of item characteristic curves for the one-parameter model are presented in Figure 2.

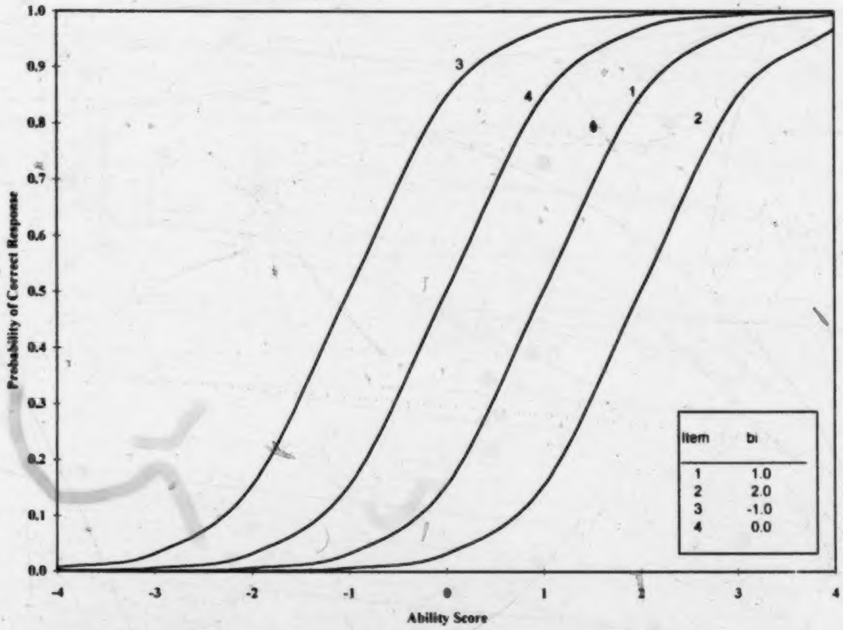


Figure 2: One-parameter Model Item Characteristic Curves for 4 Items
Redrawn from: Hambleton et al. (1991, p. 14)

The curves shown in Figure 2 differ only in their location on the ability scale since it is assumed that item difficulty is the only item characteristic that influences examinee performance. The lower asymptote for each item is zero given the assumption of no guessing.

Three-parameter Logistic Model

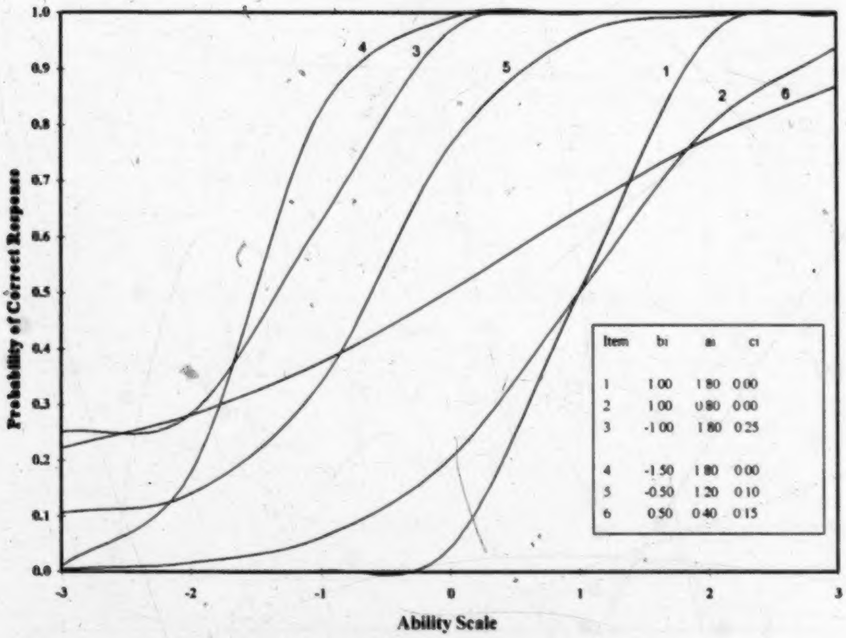
The one-parameter model is based on restrictive assumptions. It is rare to find a test with all items equally discriminating. Further, both the one- and two-parameter models make no allowance for the possibility that low-ability examinees may correctly guess an answer. Consequently, to account for guessing, a third item parameter, c_i, is added.

The guessing parameter provides a non-zero asymptote for the item characteristic curve and represents the probability that examinees of low ability will answer the item correctly. The mathematical expression for the three-parameter model is:

$$P_i(\theta) = c_i + (1 - c_i)\frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}} \qquad i = 1, 2, ..., n.$$

where b_i is the point at which the probability of correctly answering an item is $(1 + c_i)/2$.

In Figure 3, the effects of varying different parameters in the three-parameter model are illustrated. Item 4 is an example of a very easy item. Examinees with an ability level of at least 0 will answer correctly item 4, i.e., the probability of answering correctly this item is 1.00. Items 1 and 2 differ only in their discrimination parameter values. Item 1, with a higher discrimination parameter (steeper curve) more sharply discriminates among examinees than item 2 with a lower discrimination (flatter curve). The effect of guessing parameter on item characteristic curve is illustrated with items 1 and 3 which have the same item parameter values except for the guessing parameter. Since there is no possibility of guessing in item 1, there is a very small probability that examinees with low ability will answer the item correctly.



Item	b_i	a_i	c_i
1	1.00	1.80	0.00
2	1.00	0.80	0.00
3	-1.00	1.80	0.25
4	-1.50	1.80	0.00
5	-0.50	1.20	0.10
6	0.50	0.40	0.15

Figure 3: Three-parameter Item Characteristic Curves for Six Items
Redrawn from: Hambleton et al. (1991, p. 18).

Parameter Estimation

In item response models, the probability that an examinee answers an item correctly depends on the characteristics (parameters) of both the items and the person. The item and examinee parameters are estimated using maximum likelihood estimation procedure. Suppose that a randomly chosen examinee responds to a set of *n* items with response pattern $(U_1, U_2, \ldots, U_i, \ldots U_n)$. U_i is either 1 (a correct response) or 0 (an incorrect response) on item *i*. By the assumption of local independence, the joint probability of observing the response pattern is the product of the probabilities of observing each item response:

$$P(U_1,U_2,\ldots,U_n|\theta) = \prod_{i=1}^{n} P(U_i|\theta) \ .$$

Since U_j is either 1 or 0, the above equation can be written as :

$$P(U_1,U_2,\ldots,U_n|\theta) = \prod_{i=1}^{n} P_i^{U_i} Q_i^{1-U_i} \ ,$$

where $P_i = P(U_i \mid \theta)$ and $Q_i = 1 - P(U_i \mid \theta)$.

When the response pattern is observed, the expression for the joint probability is called the likelihood function and is denoted as $L(u_1, u_2, \ldots, u_i, \ldots u_n \mid \theta)$ where u_i is the observed response to item *i*. The likelihood function is therefore given by:

$$L(u_1,u_2,\ldots,u_n|\theta) = \prod_{i=1}^{n} P_i^{u_i} Q_i^{1-u_i} \ .$$

The value of θ that maximizes the likelihood function for an examinee is defined as the maximum likelihood estimate of θ for that examinee. The desired parameter estimates are those that make the likelihood function a maximum.

In practice, however, there are *m* subjects responding to *n* items leading to an *m x n* response matrix. If the *m* subjects are a random sample such that responses across subjects are mutually independent, the likelihood function for a single subject can be generalized to the joint function for the *m* subjects:

$$L(u_1,u_2,\ldots,u_n|\theta) = \prod_{j=1}^{m} \prod_{i=1}^{n} P_{ij}^{u_{ij}} Q_{ij}^{1-u_{ij}} \ .$$

The joint likelihood function is the basis for all item response theory parameter estimation processes (Suen, 1990, p. 103).

Partial Credit Model

The dichotomous ($U_i = 1$ or 0) one-parameter model has been extended to provide more information about examinee ability by awarding partial credit to each option of an item for which the options vary in terms of the degree of correctness. This extension consists of identifying one or more intermediate levels of performance on such an item and awarding partial credit for reaching one of those levels (Wright & Masters, 1982; Masters, 1988). The model is thus referred to as the partial credit model.

The partial credit model is designed for use with attitude scales, educational achievement tests, and other items that elicit responses that are scored in ordered categories. It is assumed that options can be ordered in independent steps. Each step has an associated difficulty value which indicates the probability of reaching that step along the way to successful completion of an item. The following illustration is based upon the work of Wright and Masters (1982). The probability of successfully completing a step is given by:

$$P_{jis} = \frac{e^{(\theta_j - b_{is})}}{1 + e^{(\theta_j - b_{is})}} \qquad s = 1, 2, \ldots, m_i,$$

where P_{jis} is the probability of person *j* successfully responding at the *sth* step of item *i*; b_{is} is the difficulty of the *sth* step in item *i*, and m_i is the number of steps of item *i* (Wright & Masters, 1982).

Masters (1982) presented examples of items to which the partial credit model can be applied. For a mathematics item, partial credit can be assigned for each successive step towards a solution. For example, consider the following simple numerical calculation:

$$\sqrt{\frac{7.5}{0.3} - 16} = ?$$

The steps to the solution and the credit awarded for reaching each step are:

Step	Procedure	Credit
0	Failed	0
1	7.5 / 0.3 = 25	1
2	25 - 16 = 9	2
3	√9 = 3	3

In this example there are three main steps to the solution. The steps are considered to be in a hierarchical order. The third step is awarded 3 points while the first step is awarded 1 point. The probability of a person reaching a certain step is given as the sum of the successfully completed steps prior to that step. The general expression of this probability is given by:

$$P_{xji} = \frac{e^{\sum_{s=0}^{x}(\theta_j - b_{ij})}}{\sum_{k=0}^{m_i} e^{\sum_{s=0}^{k}(\theta_j - b_{ij})}}, \quad x = 1,\ 2,\ 3, \ldots ,\ m_i ,$$

where x is the number of successfully completed steps and m_i is the total number of steps for item i. The formula gives the probability of person j scoring x on item i with m_i steps.

An example of a multiple-choice item that can be scored using the partial credit model is:

The capital city of Australia is

A. Wellington 1

B. Canberra...3

C. Montreal..0

D. Sydney...2

Aghbar and Tang (1991) applied the partial credit model when scoring cloze-type items and found that the partial credit model provided better discrimination and higher criterion related validity than did number right scoring. They concluded that "the partial credit scoring is not only more rational theoretically, but also more desirable psychometrically than the number right scoring" (p. 15).

Finite State Score Theory

Finite state score theory incorporates partial knowledge from the stand-point of a mathematical description of the variables and processes that come into play when a subject is asked to respond to multiple-choice test items. Initially developed by Hutchinson (1982), the first form of finite state theory was based upon the assumption that there are two possible states a subject can be in with regard to an item. The examinee either knows (total knowledge) or does not know (total lack of knowledge) the answer to the item. The initial form of the theory did not take into account partial knowledge. Subsequently, García-Peréz (1987) extended the theory to incorporate partial knowledge, misinformation, and guessing in addition to total knowledge and total lack of knowledge.

In its present form, finite state score theory postulates that an examinee makes independent attempts to classify every available item option as either a true or false completion of the item stem. This process determines the examinee's state of knowledge about the item and when aggregated across items, the state of knowledge of the domain tested can be determined. Two basic examinee parameters are used to determine this state. The first parameter is the ability parameter, λ ($0 \leq \lambda \leq 1$), which is defined as the proportion of all the possible statements about the subject matter whose truth value the examinee knows (García-Peréz, 1987; 1993). In this context, a statement refers to each of the options constituting the item and which has a truth value according to the content of the subject matter (García-Peréz, 1987). The lowest value, $\lambda = 0$, indicates total lack of knowledge while the highest value, $\lambda = 1$, indicates total knowledge about the subject matter. Intermediate values represent different levels of partial knowledge.

The second parameter is the guessing parameter, γ ($0 \leq \gamma \leq 1$), which represents the probability of guessing at random in the absence of assured knowledge of the correct answer (García-Peréz & Frary, 1991a). Finite state score theory also provides room for the possibility of leaving an item unanswered. The model takes into account examinees who do not know the correct answer and are not willing to guess.

The main objective of the finite state theory is to take into account the subjects' behaviour when responding to a multiple-choice item. This is the most attractive feature of the theory and reflects, at least partially, Suen's (1990) comment:

> A major criticism of existing mainstream psychometric theories is that they are essentially statistical or mathematical theories with little psychological or behavioral foundation. That is to say psychometric theories contain substantially more -metric with little psycho-. (p. 212)

Finite state theory attempts to incorporates the process used by an examinee when responding to a multiple-choice item and about the identifiability of options.

Assumptions of Finite State Score Theory

Finite state score theory is based on the following assumptions:

1. Local independence across the items. The probability of answering correctly one item is independent of the probability of answering any other item in the test;
2. Independence of options. The options within an item must be independently classifiable by examinees as if they were actually independent true-false items. Thus correct classification of fewer than all of the options must not lead the examinees to infer what the correct answer is if they do not know it;
3. All items consist of the same number of options of which one is correct and the others are incorrect;
4. Distractors are equally attractive; and
5. Conventional administration of the test. Examinees are not provided with an advice on how the test will be scored or about what guessing strategy will lead to score optimization.

States of Knowledge

For a k-option item, there are k + 1 possible states of knowledge ranging from total lack of knowledge (when no option can be classified) to total knowledge (when all options can be classified). There are k - 1 states of partial knowledge corresponding to having or not having the knowledge needed to classify 1, 2, ..., k - 1 options, regardless of whether the (single) correct option is among them.

A tree diagram can be used to facilitate the representation of how λ and γ interact to determine an examinee's response to an item (García-Peréz, 1985; García-Peréz & Frary, 1989). The tree diagram represents all possible sequences of events when an examinee attempts to respond to a multiple-choice item. To illustrate this, the tree diagram for a 3-choice item (García-Peréz, 1985) is presented in Figure 4.

The possible paths taken by an examinee when responding to a multiple-choice item are presented in Figure 4. Within a tree diagram, each node represents an event whose outcome is given by the probabilities shown beside each of the branches arising from that node. The probability of successfully classifying an option correctly is given by λ while the probability of failure to successfully classify the option is given by 1 - λ.

When reading the tree diagram (from left to right), the first, second and third nodes represent independent attempts at classifying each option as correct or incorrect. The fourth node, in the paths where they appear, represent whether the option that is the correct answer to the item is among those classified. The fifth nodes represent decisions to guess or not to guess, and the sixth nodes represent guesses when they are made.

The first path (from top to bottom) represents total knowledge, that is, knowledge to classify all three options and results in a correct answer as indicated at the end of the path. The highest degree of partial knowledge, when two of the three options are classified, is represented by paths 2, 3, and 8. This results in a correct answer.

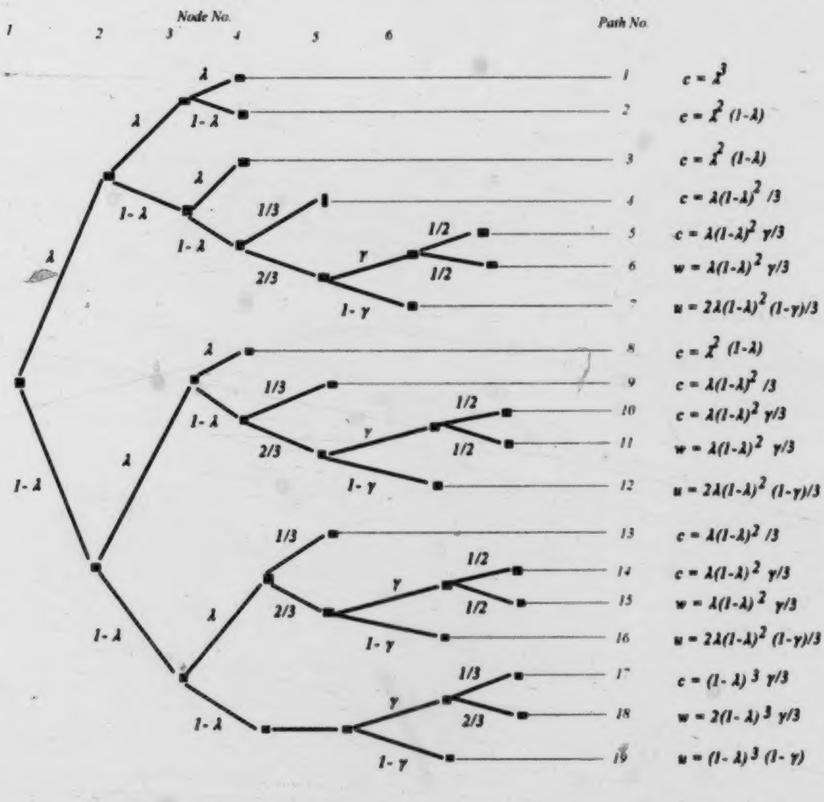

Node No.
1
2
3
4
5
6
Path No.
λ
1- λ
1/3
2/3
1/2
γ
1- γ
1 $c = \lambda^3$
2 $c = \lambda^2 (1-\lambda)$
3 $c = \lambda^2 (1-\lambda)$
4 $c = \lambda(1-\lambda)^2 /3$
5 $c = \lambda(1-\lambda)^2 \gamma/3$
6 $w = \lambda(1-\lambda)^2 \gamma/3$
7 $u = 2\lambda(1-\lambda)^2 (1-\gamma)/3$
8 $c = \lambda^2 (1-\lambda)$
9 $c = \lambda(1-\lambda)^2 /3$
10 $c = \lambda(1-\lambda)^2 \gamma/3$
11 $w = \lambda(1-\lambda)^2 \gamma/3$
12 $u = 2\lambda(1-\lambda)^2 (1-\gamma)/3$
13 $c = \lambda(1-\lambda)^2 /3$
14 $c = \lambda(1-\lambda)^2 \gamma/3$
15 $w = \lambda(1-\lambda)^2 \gamma/3$
16 $u = 2\lambda(1-\lambda)^2 (1-\gamma)/3$
17 $c = (1-\lambda)^3 \gamma/3$
18 $w = 2(1-\lambda)^3 \gamma/3$
19 $u = (1-\lambda)^3 (1-\gamma)$

Figure 4: Tree diagram representing possible sequences of events on a 3-option item. Redrawn with modications from: García-Peréz (1985, p. 65).

The next and lowest degree of partial knowledge, when one option can be classified, is represented by paths 4-7, 9-12, and 13-16. In this case, a correct answer may be given when the only classified option is the correct one (paths 4, 9, and 13) or when guessing occurs after elimination of one distractor (paths 5, 10, and 14). Guessing may also result in incorrect answer (paths 7, 12, and 16). Total lack of knowledge is represented by paths 17 to 19, with the possible outcomes of guessing and succeeding (path 17), guessing and failing (path 18) and omitting (path 19).

The probability of a path leading to either correct, incorrect or omitted response is the product of the probabilities of its several links. These probabilities are shown at the end of each path along with the class of response to the item represented by a path. Expressions for the probability of each class of response are the sum of probabilities of all paths belonging to that class. For example, the expression for the probability of answering the item correctly, c, is the sum of all paths leading to c. Therefore, from Figure 4, the probability for a correct (c), an incorrect (w), and an omitted (u) response are:

$$c = \lambda^3 + 3\lambda^2(1 - \lambda) + \lambda(1 - \lambda)^2\gamma + (1 - \lambda)^3\gamma/3$$

$$w = \lambda(1 - \lambda)^2\gamma + 2(1 - \lambda)^3\gamma/3$$

$$u = 2\lambda(1 - \lambda)^2(1 - \gamma) + 2(1 - \lambda)^3(1 - \gamma).$$

The same process for developing a 3-option tree diagram could be used to develop a 4-option diagram. However, the size of the 4-option diagram is too large for exemplification here. For the 4-option items, the probabilities of a correct, an incorrect, and omitted responses are given by the following three equations:

$$c = \lambda^4 + 4\lambda^3(1-\lambda) + 3\lambda^2(1-\lambda)^2 + 3\lambda^2(1-\lambda)^2\,\gamma/2 + \lambda(1-\lambda)^3 + \lambda(1-\lambda)^3\gamma + \lambda(1-\lambda)^4\gamma/4$$

$$w = 3\lambda^2(1-\lambda)^2\,\gamma/2 + 2\lambda(1-\lambda)^3\gamma + 3\lambda(1-\lambda)^4\gamma/4$$

$$u = 2\lambda^2(1-\lambda)^2\,(1-\gamma) + 2\lambda(1-\lambda)^3(1-\gamma) + (1-\lambda)^4(1-\gamma)$$

Several assumptions can be made about examinees' guessing behaviour depending on the instructions given to students. By incorporating these assumptions into the scoring polynomial, different scoring models can be derived to match the given instruction. For example, if the students know that the test will be scored by counting the number of correct responses, and that they will not be penalized for guessing, then it is assumed that all examinees will answer all items for which they know the answers and guess the answers for the remaining items. In this case, the guessing parameter is considered to be one ($\gamma = 1$). By replacing $\gamma = 1$, the corresponding probabilities of response outcomes can be obtained.

Another assumption about guessing behaviour, which applies in this study, is made when students are asked to answer all items without instructions as to how the test will be scored. If it can be assumed that students with partial knowledge will guess, then the willingness to guess parameter is one ($\gamma = 1$). Omission is likely to occur when an examinee lacks knowledge relevant to the item. In this case, the willingness to guess parameter is zero ($\gamma = 0$). Substituting these values in the above equations yields:

$$c = \lambda^4 + 4\lambda^3(1-\lambda) + 9/2\lambda^2(1 - \lambda)^2 + 2\lambda(1- \lambda)^3$$

$$w = 3/2\lambda^2(1 - \lambda)^2 + 2\lambda(1 - \lambda)^3$$

$$u = (1 - \lambda)^4.$$

Ability Estimation Methods

The initial method of estimating finite state scores involved taking the set of expressions for the probability of every response outcome (correct, wrong, or omitted) as a system of non-linear equations and deriving the scoring polynomial by combining these equations algebraically. The probability of a response outcome was replaced by the empirical proportion of items answered by an examinee in a corresponding category. More specifically, the method involved finding the root, the value of λ that satisfied the equation, of the scoring polynomial in the interval [0,1].

Ability estimates based on the root of the scoring polynomial lack statistical properties. Following this drawback, García-Peréz (1994) discontinued the use of this method in favour of statistical estimation procedures. In his 1994 study, he compared several parameter estimation methods and goodness-of-fit statistics in an attempt to develop guidelines on how to choose one method over others. His findings led him to conclude that "it is virtually impossible to advance any general recommendation as to what parameter estimation method or goodness-of-fit statistics to use" (p. 277). Subsequently, García-Peréz (February 7, 1995, personal communication) suggested the use of the minimum chi-square estimation method. The minimum chi-square statistic is given by:

$$\chi^2 = \sum \frac{[n_k - nP_k(\lambda)]^2}{nP_k(\lambda)} ,$$

where *k* represents the categories (correct, incorrect, and unanswered), *n* is the number of items and *P* is the probability of the response outcome. Note that *P* indicates the theoretical probability of each response outcome, that is c, w, and u. The complete chi-square formula, incorporating the values for c, w, and u is:

$$\chi^2 = \frac{[n_c - n_c(\lambda^4 + 4\lambda^3(1-\lambda) + 9/2\lambda^2(1-\lambda)^2 + 2\lambda(1-\lambda)^3)]^2}{n_c[\lambda^4 + 4\lambda^3(1-\lambda) + 9/2\lambda^2(1-\lambda)^2 + 2\lambda(1-\lambda)^3]}$$

$$+ \frac{[n_w - n_w(3/2\lambda^2(1-\lambda)^2 + 2\lambda(1-\lambda)^3)]^2}{n_w[3/2\lambda^2(1-\lambda)^2 + 2\lambda(1-\lambda)^3]}$$

$$+ \frac{[n_u - n_u(1-\lambda)^4]^2}{n_u(1-\lambda)^4} .$$

Studies Involving Finite State Score Theory

García-Peréz and Frary (1989) conducted a simulation study to evaluate the psychometric properties of finite state scores based on different assumptions about guessing. They also examined the extent to which these scores provide similar ranking of examinees. They found correlations of .99 between scores computed using the finite state number-right scoring assumption and scores based on the finite state formula scoring assumption. This finding suggests that the assumptions about examinees guessing behaviour do not lead into differences in the ranking of examinees using the ability estimates.

García-Peréz (1993) applied the finite state score theory to refute the claim that items with the last option reading "none of the above" (NOTA items) have no advantage over conventional items, that is, items whose options are content based. He used maximum likelihood estimation procedure to obtain point and interval estimates for the ability parameter λ. The differential efficiency of NOTA and conventional items was assessed in terms of the accuracy of the point estimates and the width of the confidence interval obtained for each type of the items. Simulated data were used to facilitate this purpose. Residual analysis for both type of items was done to determine the fit of data to the model. The findings also revealed that NOTA items yielded narrower confidence intervals than their conventional counterparts. García-Peréz (1993) found that for $\lambda \geq .7$, examinees' abilities were overestimated from their responses to conventional items, yielding positive mean residuals. Based on these findings, García-Peréz (1993) concluded that NOTA items are preferable to conventional items when accuracy in ability estimation is a goal.

Zin (1992) examined the extent to which finite state scores and conventional number-right scores rank order examinees similarly. She also examined the effect of varying the guessing assumptions. Responses of 5,268 eleventh grade students to the science subtest of Test of Achievement and Proficiency (Riverside Publishing Company, 1986, cited in Zin, 1992) were analyzed in her study. This subtest contained 54 multiple-choice items with 4 options. The ability estimates were obtained through finding the root of the scoring polynomial she derived. Zin (1992) obtained Spearman rank order correlations of .998 between the scores. Furthermore, she found that differing assumptions about guessing strategy had little effect on the magnitude of the ability estimates. This finding was consistent with the earlier finding of García-Peréz and Frary (1989).

Ndalichako and Rogers (1997) compared number right, item response (one-, two-, and three-parameter ability estimates), and finite state scores. Responses of 1,232 subjects on 70 multiple-choice items contained in a provincially administered school-leaving examination of Reading Comprehension (Alberta Education, 1992) were analyzed. The correlations between the ability estimates were very high (≥ .96). The mean absolute differences among the pairs of transformed scores ranged from .77 to 2.15. The closest agreement were between one parameter and the two-parameter item response ability estimates (.77). The least agreement were between the number right and finite state score ability estimates (2.15).

Concluding Remarks

In spite of the theoretical promise shown by finite state score theory in overcoming the inherent weaknesses in scoring multiple-choice test items, there is still not enough evidence to draw definitive conclusions about the validity of inferences drawn from the ability estimates, λ. Of the four comparative studies in which the finite state score theory was used, simulated data were used in two. More comprehensive studies using "real" as opposed to simulated data are required before more definitive conclusions can be made about the utility of finite state score theory.

The intent of the present study was to provide more information about the utility of finite state score theory in scoring multiple-choice test items in an actual field setting. Further, the present study was an extension of the study by Ndalichako and Rogers' (1997. While Ndalichako and Rogers (1997) analyzed all the items in the English examination, in the present study two subtests were considered in terms of adherence to the assumption of independence among the item options. To further examine the effect on ability estimates of violating the assumption of independence among the item options, two subtests were formed from a Test of Test-Wiseneness.

CHAPTER 3: METHOD

The purpose of this study was to apply finite state score theory in scoring multiple-choice items and to assess the results by comparing them to the results obtained using number right and item response scoring approaches when the assumption of option independence was violated. This study was designed to answer the two specific research questions presented in Chapter 1 and restated here for information:

1. To what extent are ability estimates from finite state scores, number right, and item response models similar to each other when the assumption of independence of options made in finite state score theory is violated?
2. To what extent do the conventional and item response item analyses lead to the same decisions regarding item quality?

The method used to address these two questions is described in this chapter. First, the data that were used are described. This is then followed by a description of the analyses that were undertaken and the computer programs that were used to complete these analyses.

Data Sources

Two sets of data were analyzed. The first set consisted of responses of 1,232 students to the June, 1992 form of English 30 Diploma Examination (Alberta Education, 1992) administered in the province of Alberta, Canada. This is a provincial school leaving examination which contributes 50% toward a student's final course grade. The students whose responses were analyzed in the present study were those who participated in a study of the influence of test-wiseness upon student performance. In addition to the Diploma Examination, these students also completed a Test of Test-wiseness (Rogers & Wilson, 1993). Test-wiseness is a cognitive ability or set of skills which a test taker can use to improve a test score (Sarnacki, 1979). The same students' responses to the Test of Test-wiseness formed the second data set.

English 30 Diploma Examination: Reading Section.

The English 30 Diploma Examination consists of two sections, a written response section and a reading section. The reading section, which contained 70 four-option multiple-choice items, is the section that was analyzed in this study. Each item was classified by Alberta Education (1992) in two ways: according to the curricular content area being tested and according to the thinking skills demanded by the item. The distribution of items according to these classifications is presented in Table 1.

Table 1

Table of Specifications for English 30 Diploma Examination, 1992.

Course Content	Thinking Skills		
	Literal Understanding	Inference and Application	Evaluation
Meanings	10, 11, 13, 15, 67	1, 4, 19, 20, 22, 23, 24, 25, 28, 41, 42, 44, 45, 49, 50	5, 12, 14, 16, 18, 29, 36, 47, 51, 52
Critical Response	59	3, 6, 7, 9, 17, 27, 33, 35, 40, 43, 48, 55, 56, 57, 58, 61, 63, 69	8, 32, 46, 53, 54, 60, 62
Human Experience and Values	26	2, 30, 34, 37, 38, 39, 64, 65, 66, 68	21, 31, 70

Source: Alberta Education (1992)

The reading section of the English 30 Examination contained two subtests of items that were of interest in this study. The first subtest consisted of 48 items which satisfied the assumption of option independence. The second subtest consisted of 20 items that required examinees to provide the "best or the most correct" answer, a condition which, if present, negates the assumption of independence among the options. Two items, each containing a pair of opposite options (Rogers & Wilson, 1993), were excluded from the analyses to avoid confounding the interpretation of any influence of dependence of options in the first subtest and to avoid mixing "best answer" and "correct answer" items in the second subtest.

Test of Test-Wiseness.

The Test of Test-Wiseness measures the degree to which students use deductive reasoning strategies to answer multiple-choice test items. Four test-wise elements, identified in the previous provincial government examinations, administered in Alberta and British Columbia, were assessed by the Test of Test-Wiseness. The definition of these elements, adopted from Millman (1966), are:

- ID1 - eliminate options known to be incorrect (absurd options);
- ID2 - choose neither or both of two options which imply the correctness of each other (similar options);
- ID3 - choose neither or one (but not both) of two options, one of which, if correct, would imply the incorrectness of the other (opposite options);
- IIB4 - recognize and use similarities between the stem and the options (cued options).

Options in items with ID2 and ID3 elements are not independent of each other. García-Peréz and Frary (1989) indicated that "items to be scored using finite state score theory should not have two options where one is simply the negation or the opposite of the other" (p. 408). Items with ID2 options, on the other hand, violate the assumption of the independence of options since the similarity between options implies that in a single-correct-answer item, neither can be correct.

The Test of Test-Wiseness contained 50 four-option multiple-choice items. These items were classified in terms of content familiarity (to the students) and susceptibility to a test-wise deductive reasoning element. This classification of the items is summarized in Table 2. Items classified as novel contained subject content which was unfamiliar to students. The content based items included items selected from school leaving examinations administered in Alberta and British Columbia.

Table 2:

Classification of the Test of Test-wiseness Items

Test-wiseness Principle	Content Area				
	English	Mathematics	Social Studies	Biology	Total
Novel					
ID1	2	1	2	1	6
ID2	2	1	1	2	6
ID3	2	2	-	1	5
IIB4	1	1	3	2	7
Sub-total	7	5	6	6	24
Content Based					
ID1	-	2	2	2	6
ID2	-	-	1	2	3
ID3	-	1	1	2	4
IIB4	-	2	3	2	7
Sub-total	-	5	7	8	20
No test-wise cue	-	2	2	2	6
Total	7	12	15	16	50

Source: Rogers and Wilson (1993).

Like the English 30 Examination, the test of Test-Wiseness contained two subtests that were of interest in the present study. The first subtest consisted of 32 items which satisfied the option independence assumption. The second subtest contained 18 items that included either a pair of similar options or a pair of opposite options. An example of items that included a pair of similar options (A and D) is:

Mr. Adams, in Henry Fielding's Joseph Andrews,

A. learns his parents were of the nobility.

B. takes sick after falling through the ice.

C. falls into the mud while reading.

D. discovers he is of noble birth.

The next example illustrates an item with a pair of opposite options (A and B):

Compared to normal cells, bileuvial cells

A. divide more rapidly.

B. divide more slowly.

C. have more cytoplasm.

D. have more mitochondria.

Data Analyses

The psychometric characteristics of the items within each of the two subtests of English 30 Reading Examination and the two subtests of Test of Test-Wiseness were examined using conventional analysis procedures and three or four item response models depending on the nature of the items.

Conventional Item Analysis

Conventional item analysis is rooted in classical test theory. In the context of school leaving examinations, such as the one considered in the present study, the aim is to identify items which when taken together as a set, maximize the reliability of scores derived from the item set. At the same time, attention must be directed toward ensuring that the scores can be validly interpreted with reference to the domain assessed (Cronbach, 1971, p. 458).

In a conventional item analysis, the assessment of multiple-choice item involves assessment of the correct option and each of the distractors in terms of the number and kinds of students who respond to each option. The item's difficulty is equal to the proportion of examinees who select the correct option. The item's discimination indicates how well the item discriminates between high-achieving and low-achieving examinees. The selection of the discrimination index depends upon the interpretations to be made from the test scores. In the case of norm-referenced score interpretation, several indices are often considered, including the point-biserial correlation, corrected point-biserial correlation, biserial correlation, and the discrimination (D) index. Most measurement specialists appear to favour the uncorrected point-biserial (e.g., Ary, et al., 1996; Haladyana, 1994; Crocker & Algina, 1986; Sax, 1995). Alberta Education uses the corrected point-biserial for the correct option and the uncorrected point-biserial for the distractors.

Given that the responses on the two subtests of the English 30 Examination are a subset of responses collected by Alberta Education, the discrimination indices used by Alberta Education were adopted for the present study. That is, the corrected point-biserial was used for the correct answer and the uncorrected point-biserial was used for the distractors. However, both corrected and uncorrected point-biserial correlations for the correct option were computed because the uncorrected coefficient is often used in item response theory. The uncorrected point-biserials were used for the items in the two subtests of the Test of Test-Wiseness because the developers of this test used only this coefficient (Rogers & Wilson, 1993).

The evaluation criteria for the p-values and item discrimination indices used by Alberta Education were employed in the present study. These criteria include:

(a) Minimum and maximum acceptable difficulty levels of, respectively, .30 and .85;

(b) Minimum acceptable corrected point-biserial of .20 (the same value was used for the uncorrected point-biserial);

(c) Negative point-biserials for the distractors; and

(d) Difficulty levels of at least .05 for the distractors (Alberta Education, 1992).

The evaluation criteria used by Alberta Education for the correct option (a and b) are more specific than the evaluation criteria for the distractors (c and d). As Haladyana (1994) noted, "textbooks seldom give an in depth treatment of this subject, probably because not much is known about how to evaluate distractors" (p. 153).

Item Response Item Analysis

The success of the application of an item response model is based upon the appropriateness of the model for the data under consideration. In this study, the assumptions underlying the use of item response models considered were tested prior to their use. The following item response assumptions were examined: unidimensionality, equal discrimination indices, non-speeded test administration, and guessing. An additional assumption concerns the presence of local independence. As explained earlier, this assumption is met when the item responses are unidimensional or when all the dimensions influencing the examinees' performances are accounted for.

Unidimensionality. The assumption of unidimensionality underlies all the item response models considered. This assumption was assessed using three methods. The first method involved the use of principal component factor analysis using both phi correlations and tetrachoric correlations in order to examine the convergence of the results. The SPSS 6.1 computer program (SPSS Inc., 1994) was used to complete the analyses.

The most commonly used method for assessing dimensionality of item responses is the amount of variance accounted for by the first component compared to the others. The expectations were that (1) the first component would be quite large thereby accounting for a large proportion of variance and (2) the difference between the first and second components would be large while the differences between successive pairs of adjacent components, beginning with component 2, would be negligible. Often, however, the results are not as clear as expected. Hambleton et al. (1991) found that what is required for the unidimensionality assumption to be met is the presence of a dominant component or factor that influences the test performance" (p. 9). Later, Huynh and Ferrara (1994) observed that "good ability estimates can be obtained even if the first component accounts for less than ten percent of the total test variance" (p. 127).

The second method involved the application of the scree test (Cattell, 1952). This test was used to clarify and confirm the results from the first method. The third method involved the use of Stout's T statistic (Nandakumar & Stout, 1993; Stout, 1987, 1990). Stout's T statistic has been shown to discriminate well between essentially unidimensional and multidimensional sets of test scores for both simulated (Nandakumar, 1994; Nandakumar & Stout, 1993; Stout, 1987) and real data (Nandakumar, 1993, 1994). The hypotheses that are tested using the Stout's T statistic are:

$$H_0: d_E = 1 \text{ versus } H_1: d_E > 1,$$

where d_E denotes the essential dimensionality of the item responses. To compute Stout's T statistic, the test is divided into two subtests, say AT1 and AT2 formed on the basis that the items within them appear to represent a dominant ability. Then the following sequence of formulas are used to determine the value of Stout's T statistic:

$$Y_{jk} = \sum_{i=1}^{n} U_{ijk} / n \ ,$$

$$\bar{Y}_k = \sum_{i=1}^{J_k} Y_{jk} \ / \ J_k \ ,$$

$$\hat{p}_{ik} = \sum_{j=1}^{J_k} U_{ijk} \ / \ J_k \ ,$$

where U_{ijk} (1 or 0) denotes the response for item i, $i = 1, 2, \ldots, n$, by examinee j in subgroup k, and J_k denotes the total number of examinees in subgroup k, $j = 1, 2, \ldots, J_k$. The variance of the subtest score for subgroup k and the "unidimensional" variance estimate for subgroup k are, respectively:

$$\hat{\sigma}^2_{k} = \sum_{j=1}^{J_k} (Y_{jk} - \bar{Y}_k)^2 / J_k ,$$

and

$$\hat{\sigma}^2_{U,k} = \sum_{i=1}^{n} \hat{p}_{ik}(1 - \hat{p}_{ik}) \,/\, n^2 .$$

Finally,

$$T_i = \frac{1}{K^{1/2}} \sum_{k=1}^{K} \left[\frac{\hat{\sigma}^2_{ik} - \hat{\sigma}^2_{U,ik}}{S_{ik}} \right], i = 1 \text{ or } 2.$$

and

$$T = \frac{(T_1 - T_2)}{\sqrt{2}} ,$$

where S_k is a normalizing constant for subgroup k.

The DIMTEST computer program, which was initially developed by Stout (1987, 1990) and subsequently refined by Nandakumar and Stout (1993), was used to compute Stout's T statistic. An automatic execution procedure was employed. In this procedure, the subtests AT1 and AT2 are automatically formed by the program using the factor loadings of the items. The value of Stout's T statistic is interpreted in terms of the unit normal distribution. The *p* values for essentially unidimensional tests are expected to be large whereas the *p* values for multidimensional tests are expected to be less than or equal to the specified level of significance (Nandakumar, 1994, p. 23).

The expectation was that the proportion of variance rule, the scree test, and Stout's T statistic would lead to the same decision about the dimensionality of item responses. However, as will be shown, these statistics did not point to the same conclusion for the subtest of 18 Test of Test-Wiseness items that failed to meet the option independence assumption. The decision was taken at that point to proceed with the item response item analyses and examine the fit of items to the corresponding item response models.

Discrimination. The assumption of equal item discrimination is made in the case of the one-parameter model. In contrast, the two- and three-parameter models do not require this assumption. Hambleton and Murray (1983) suggested that to satisfy the condition of homogeneity of item discrimination, the difference between the lowest and highest uncorrected point-biserial correlations should be less than .15.

Speededness. To determine whether or not the examination was speeded, the percentage of examinees who did not complete the test was examined. Hambleton et. al. (1991) recommended this approach and suggested that the percentage of examinees completing the test, percentage of examinees completing 75% of the test, and the number of items completed by 80% of the examinees be reviewed. "When nearly all examinees complete nearly all of the items, speed is assumed to be an unimportant factor in test performance" (p. 57). In the present study, for this assumption to be satisfied, the percentage of examinees who completed the last three items in a test was examined. If 95% of examinees completed these items then it was considered that speed was not a factor that influenced examinees performances.

Guessing. Both the one- and two-parameter models do not include a guessing parameter. Thus, it is an assumption of these two models that guessing has minimal influence on examineesperformances. This assumption was assessed by examining the performance of low-achieving examinees (based on their total score). In this study, low-achieving examinees were operationally defined as examinees whose number right score on a subtest under consideration was less than one third of the total score. The expectation was that these examinees would not be able to give correct answers to the most difficult items. Three items with the lowest p-values were considered the most difficult items. Consequently, close to zero performance of low-scoring examinees on such items would support the viability of the assumption that guessing was minimal.

Examination of the Fit of Item Response Models

To test the extent to which the responses for each item fit each of the item response models, different procedures were used for the one-, two-, and three-parameter item response models and the partial credit item response model. These procedures are outlined below.

One-, two,- and three-parameter models. The method of examining the fit of item responses to the one-, two-, and three-parameter models depends on the number of items in the test analyzed. For short tests (11 to 19 items), the root mean square standardized residuals $RMS(\delta_{ik})$ is used. This is computed using the following equation:

$$RMS(\delta_i) = \left[\sum_{k}^{q} \bar{N}_k \delta^2{}_{ik} \,/\, \sum_{k}^{q} \bar{N}_k \right]^{\frac{1}{2}},$$

where $\bar{N}_k$ is the expected number of attempts at ability point X_k, and δ_{ik} is the standardized difference between the probability of a correct response for item i at selected values of θ and the probabilities at ability point X_k computed from the corresponding fitted response model. δ_{ik} is computed as:

$$\delta_{ik} = \frac{\sum_{l}^{S} W_{lk} \, [x_{li} - P_i(X_k)]}{\sqrt{\sum_{l}^{S} W_{lk} \, [x_{li} - P_i(X_k)]^2}} ,$$

where

$$W_{lk} = \frac{f_l P(X_l \mid X_k)}{P(X_l)} ,$$

$P_i(X_k)$ is the probability of answering item i correctly at ability X_k, and f_l is the observed frequency of response pattern l. RMS values greater than 2 are taken as an indication that item responses fail to fit the model satisfactorily (Mislevy & Bock, 1990).

For long tests (20 items or more) the chi-square statistic, is used to assess the fit of item response data to the models:

$$\chi^2_i = 2\sum_{h=1}^{n_g} \left[f_{hi} \log_e \frac{f_{hj}}{N_h P_i(\bar{\theta}_h)} + (N_h - f_{hi}) \log_e \frac{N_h - f_{hj}}{N_h[1 - P_i(\bar{\theta}_h)]} \right] ,$$

where n_g is the number of intervals on the θ-continuum that are formed on the basis of examinees' estimated value of θ, f_{hi} is the observed frequency of correct responses to item i in interval h, N_h is the number of examinees assigned to that interval, and $P_i(\bar{\theta}_h)$ is the value of the fitted response function for item i at $\bar{\theta}_h$, the average ability of respondents in interval h (Mislevy & Bock, 1990). A significant chi-square ($p \leq .05$) indicates a lack of fit of the item response data to the theoretical model (Hambleton, et al., 1991; Mislevy & Bock, 1990).

The RMS statistic was used to assess model fit for the subtest of 18 Test of Test-Wiseness items that did not satisfy the option independence assumption. The chi-square statistic was used to assess model fit for the item response data for three subtests: the two subtests formed from the English 30 Examination and the subtest of 32 Test of Test-Wiseness items that satisfied the assumption of option independence.

Partial credit model. The fit of items to the partial credit model, which was applied to the 20 "best answer" items contained in the English 30 Examination, was tested using the infit and the outfit statistics (Wright & Linacre, 1992). The infit statistic is a standardized information-weighted mean square statistic. It is more sensitive to unexpected behaviour affecting responses to items near an examinee's ability level. The formula for the infit statistic is:

$$v_i = \frac{\sum_{j=1}^{N} W_{ji} Z^2_{ji}}{\sum_{j=1}^{N} W_{ji}} ,$$

where v_i is the weighted mean square, W_{ji} is the variance of the observed response for person j on item i, Z^2_{ji} is the squared standardized residual between the expected value of the response for each person and the observed response, and N is the number of people.

The outfit statistic is a standardized outlier-sensitive mean square fit statistic. It is more sensitive to unexpected behaviour by examinees on items far from their ability (Wright & Linacre, 1992). The formula for the outfit statistic is:

$$u_i = \frac{\sum_{j=1}^{N} Z^2_{ji}}{N} ,$$

where u_i is the unweighted mean square.

Items whose infit or outfit statistics were greater than 3 were considered as misfitting the partial credit model.

Ability Estimation

Abilities were estimated using the number right scores, and estimates produced by the item response and finite state score models. The number right scores were not corrected for chance. Marginal maximum likelihood estimation procedure was used to obtain the ability estimates for the one-, two-, and three-parameter item response models while the unconditional maximum likelihood estimation procedure was used in the case of partial credit model. The minimum chi-square estimation procedure was used to obtain the finite state ability estimates.

Computer programs. The ITEMAN computer program (Assessment System Corporation, 1992) was used to obtain the number right scores and to complete the conventional item analysis. In the case of the one-, two-, and three-parameter item response models, ability and item parameter estimates were obtained using the BILOG computer program (Mislevy & Bock, 1990); in the case of the partial credit model, ability estimates and the difficulty parameter were obtained using the BIGSTEPS computer program (Wright & Linacre, 1992). Lastly, a FORTRAN program adopted from García-Peréz (1994) and translated into SPSS language (see Appendix A) was used to obtain the finite state scores ability estimates.

Comparison of Misfitting Items

The items identified as misfitting by the conventional item analysis and by the item analyses corresponding to each of the item response models considered were compared. The purpose of the comparison was to examine the extent to which conventional item analysis and item response item analyses led to the same decision regarding the item quality. These analyses were completed separately for each of the four subtests.

Agreement Among Ability Estimates

Ability estimates from finite state score theory were compared with number right scores and with ability estimates from item response theory for each of the four subtests. The purpose of these comparisons was to examine whether or not the finite state scores provided information similar to the score information provided by the number right scores and item response models when the assumption of option independence was violated.

The comparisons were made using two different procedures. First, to examine if the students were ranked differently by the different scoring systems, correlations among the ability estimates fromfinite state score model and those from each of the item response models and the number right scores were computed and compared.

Scatterplots were examined to check for linearlity. The graphs for all pairs of estimates except those involving the finite state scores were linear. In the case of the pairs in which the finite state scores were involved, there was some evidence of a curvilinear relationship. However, the test of nonlinearlity (Glass & Hopkins, 1984, p. 318) revealed that the departures from linearlity was not significant at the .05 level of significance. Consequently, the Pearson-product moment correlation was computed and examined for all pairs of the ability estimates for each subtest.

Second, to examine if the scores for a student were equal in an absolute sense, the mean absolute difference among pairs of ability estimates were examined. Since the scores were expressed in different metrics, they were first converted to T-scores ($\mu = 50$, $\sigma = 10$) (Glass & Hopkins, 1984). The T-scores were then used to compute the mean absolute differences among the estimates. The extent to which the models provided similar or different information was examined using the mean absolute deviation (MAD) between the transformed scores and the standard deviations (S) of the absolute scores. The equations used were:

$$MAD_{xy} = \frac{\sum_{j=1}^{N} |T_{xj} - T_{yj}|}{N},$$

and

$$S = \sqrt{\frac{\sum_{j=i}^{N} (T_j - \bar{T})^2}{N}},$$

where T_x and T_y are the corresponding transformed scores yielded by models x and y; and N is the number of subjects.

CHAPTER 4: RESULTS BASED ON THE ENGLISH 30 EXAMINATION

Results of the analyses based on the English 30 Examination are presented in this chapter; the results based on the analyses of the Test of Test-Wiseness subtests are presented in the next chapter. The results and findings presented in the two chapters are organized in the same way. Further, the text is somewhat repetitious. This approach was taken to facilitate the comparison of results across subtests. As explained in Chapter 3, two subtests were formed from the English 30 Examination. The first subtest contained 48 items which satisfied the option independence assumption. The second subtest contained 20 items which required examinees to provide the best answer from the given item options and so did not satisfy the option independence assumption. Likewise, two subtests were formed from the Test of Test-Wiseness.

For the conventional and item response analyses procedures, the analyses and comparisons were completed at the item and subtest levels. In the case of finite state score theory, analyses at the item level are not possible; consequently, the analyses and the subsequent comparisons were completed at the subtest level only.

Each chapter is organized into three main sections. The first two sections contain results for the two subtests formed from the English 30 Examination and from the Test of Test-Wiseness. Within each section a description of subtest characteristics is provided first. This is then followed by the results of the conventional and item response item analyses. The results at the subtest level are then presented and discussed in terms of the correlations among the subtest score estimates, mean absolute deviations and standard deviations of the absolute differences among all pairs of ability estimates. In the last section, differences among the correlations, means and standard deviations of the absolute differences among the ability estimates are summarized and discussed in terms of the item analyses and other factors identified during the analyses.

English 30 Items Which Satisfied the Option Independence Assumption

The results presented in the following section are for the 48 items in the English 30 Reading Comprehension Examination that satisfied the assumption of option independence. These items required examinees to select the correct answer from the four given options. For convenience, this subtest is referred to as ENGI, where I denotes independent options.

Subtest Characteristics

The distribution of number right scores for the ENGI subtest is presented in Figure 5. The mean score was 33.52 (69.8 %) and the corresponding standard deviation was 7.57 (15.8 %). The internal consistency index (Cronbach's alpha) was .85 and the standard error of measurement was 2.91 (6.0%). The value for Cronbach's alpha indicates a high degree of internal consistency or item homogeneity. The skewness coefficient was -.34 and the kurtosis coefficient was -.44. Together, these values indicate that the distribution of ENGI scores was skewed to the left and was platykurtic or flatter than the normal curve.

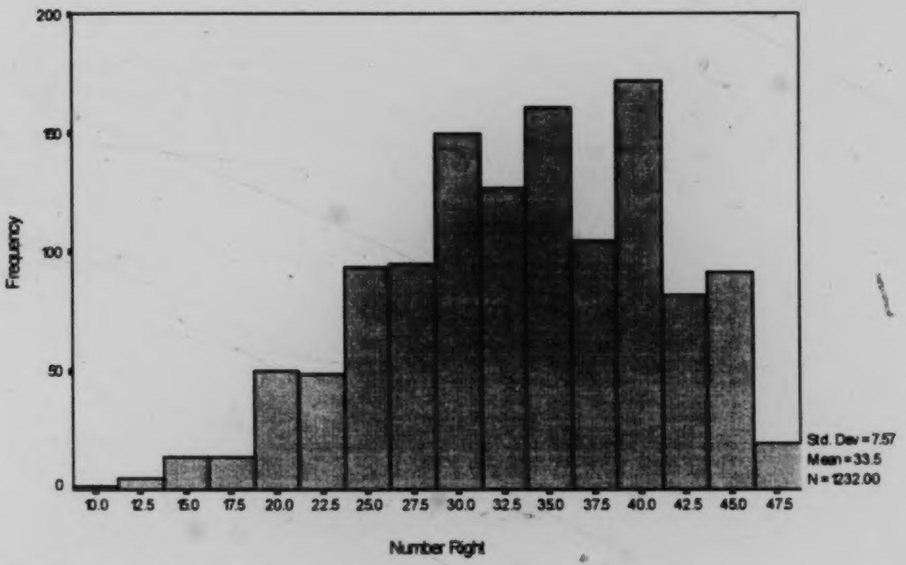


Figure 5: Distribution of Number Right Scores: ENGI Subtest

Item Characteristics

The estimated item parameters obtained from the conventional and item response item analyses are listed in Table 3 for each of the 48 items in the ENGI subtest. The results are presented together to facilitate comparisons. The conventional analyses include difficulty index (p-value), both the corrected (cpbs) and uncorrected point-biserial (pbs) correlations. Furthermore, the incorrect options whose p-values did not meet the minimum value of .05 are listed. For the item response item analyses, difficulty indices (b) for the one-, two-, and three-parameter models are presented whereas the discrimination indices (a) for the two- and three-parameter models and guessing parameter (c) for the three-parameter model are provided. Misfitting items are identified by an asterisk (*).

Table 3

Conventional and Item Response Item Analysis Results: ENGI Subtest

Item	Conventional				IRT Parameter Estimates					
					1PL	2PL		3PL		
	p-value	cpbs	pbs	option	b	b	a	b	a	c
1	0.84	0.16*	0.21	a, c	-2.36	-3.16	0.33	-2.37	0.35	0.26
2	0.79	0.30	0.34	a	-1.83	-1.66	0.54	-0.88	0.66	0.31
3	0.66	0.40	0.45	✓	-0.95*	-0.76	0.66	-0.19	0.89	0.25
4	0.72	0.33	0.39	✓	-1.33	-1.18	0.56	-0.61	0.66	0.24
6	0.60	0.21	0.28	b	-0.58*	-0.79	0.33	0.84	0.80	0.43
7	0.75	0.22	0.27	b	-1.54	-1.92	0.36	-0.94	0.43	0.28
9	0.51	0.24	0.31	c	-0.03*	-0.06	0.37	0.98	0.91	0.34
10	0.52	0.28	0.34	b	-0.11	-0.15	0.39	0.44	0.49	0.18
11	0.78	0.41	0.46	b, d	-1.78*	-1.23	0.80	-0.86	0.93	0.21
13	0.75	0.38	0.43	a	-1.54*	-1.16*	0.70	-0.56	0.91	0.28
15	0.83	0.25	0.30	b	-2.17	-2.09	0.50	-1.44	0.55	0.27
17	0.77	0.29	0.34	b	-1.72	-1.60	0.52	-0.86	0.62	0.29
19	0.82	0.27	0.31	b	-2.11	-1.99	0.50	-1.35	0.56	0.27
20	0.68	0.30	0.35	✓	-1.07	-1.07	0.47	-0.46	0.56	0.22
23	0.79	0.33	0.37	a	-1.87	-1.55	0.61	-1.14	0.66	0.21
24	0.75	0.41	0.46	✓	-1.53*	-1.08	0.79	-0.69	0.91	0.21
25	0.40	0.36	0.42	✓	0.62*	0.52*	0.57	0.82*	0.85	0.15
26	0.50	0.27	0.33	✓	0.03	0.01	0.40	0.75	0.62	0.23
27	0.46	0.41	0.46	✓	0.25*	0.17*	0.67	0.60	1.47	0.22
28	0.63	0.38	0.44	c	-0.71*	-0.56	0.62	0.94	0.79	0.21
30	0.55	0.22	0.28	✓	-0.24*	-0.34	0.33	-1.12	0.66	0.35
33	0.76	0.21	0.27	✓	-1.65	-2.08	0.36	-0. 36	0.41	0.28
34	0.65	0.35	0.41	✓	-0.89*	-0.80	0.55	-0.36	0.65	0.18
35	0.76	0.19*	0.25	b	-1.62	-2.17	0.33	-0.99	0.40	0.31
37	0.81	0.27	0.31	a	-2.06	-2.00	0.49	-1.42	0.53	0.25

Table 3 Continued

Item	Conventional				IRT Parameter Estimates					
					1PL	2PL		3PL		
Item	p-value	cpbs	pbs	option	b	b	a	b	a	c
38	0.79	0.25	0.30	✓	-1.84	-1.91	0.45	-1.02	0.54	0.31
39	0.84	0.31	0.35	✓	-2.37*	-1.91	0.62	-1.48	0.68	0.23
40	0.60	0.36	0.41	✓	-0.57	-0.53	0.54	-0.05	0.67	0.19
41	0.91*	0.29	0.32	a, c	-3.17*	-2.24	0.75	-1.96	0.76	0.24
42	0.58	0.35	0.40	✓	-0.47*	-0.44	0.54	0.03	0.67	0.18
44	0.66	0.37	0.43	✓	-0.95*	-0.79*	0.62	-0.12	0.89	0.28
45	0.72	0.38	0.43	b	-1.35*	-1.04*	0.68	-0.23	1.09	0.36
48	0.56	0.27	0.33	✓	-0.31	-0.37	0.40	0.43	0.58	0.25
49	0.76	0.28	0.33	✓	-1.65	-1.66	0.47	-1.09	0.52	0.22
50	0.82	0.37	0.41	✓	-2.12*	-1.49	0.78	-0.89	0.97	0.32
55	0.51	0.32	0.37	d	-0.05*	-0.07*	0.48	0.75	1.15	0.32
56	0.78	0.27	0.32	c	-1.80	-1.81	0.47	-1.28	0.51	0.22
57	0.74	0.33	0.39	✓	-1.45	-1.26	0.57	-0.61	0.70	0.27
58	0.73	0.13*	0.18*	d	-1.39*	-2.56	0.23	-0.95	0.28	0.31
59	0.56	0.28	0.34	✓	-0.32	-0.38	0.40	0.53	0.64	0.28
61	0.70	0.33	0.38	✓	-1.20	-1.07	0.55	-0.34	0.73	0.29
63	0.75	0.29	0.35	a	-1.52	-1.45	0.50	-0.51	0.68	0.34
64	0.72	0.41	0.45	✓	-1.35*	-1.00	0.73	-0.49	0.92	0.25
65	0.84	0.24	0.29	c, d	-2.31	-2.33	0.46	-1.81	0.49	0.24
66	0.69	0.33	0.38	✓	-1.09	-1.03	0.52	0.40	0.64	0.24
67	0.77	0.46	0.51	✓	-1.69*	-1.06	0.99	-0.65	1.26	0.24
68	0.63	0.22	0.28	b	-0.76	-1.02	0.33	0.20	0.48	0.31
69	0.75	0.27	0.33	✓	-1.54	-1.53	0.48	-0.46	0.66	0.36

Conventional Item Analysis Results

Results yielded by the conventional item analysis are presented in columns 2 to 5 of Table 3. As shown in column 2, the item difficulties for the correct options ranged from .40 to .91. Adopting the evaluation criteria used by Alberta Education, all items met the minimum difficulty standard of .30. One item, item 41 (p-value = .91) exceeded the maximum difficulty standard of .85. Thus, one item failed to meet the first standard for the difficulty of an item.

The corrected point-biserial item discrimination indices for the correct options ranged from .13 to .46. Three items, 1 ($_{c}r_{pbs} = .16$), 35 ($_{c}r_{pbs} = .19$), and 58 ($_{c}r_{pbs} = .13$), failed to meet the minimum standard of .20. Thus, at this point, four items, 1, 35, 41, and 58, failed to meet at least one of the two standards set for the correct or key option.

As mentioned previously, the uncorrected point-biserial correlation is used in the assessment of the assumption of equal discrimination in the case of the one-parameter and partial credit models. The values for this coefficient, presented in the fourth column of Table 3, ranged from .18 to .51. Adopting this coefficient, only item 58 failed to meet the minimum standard of .20.

All 48 items met the third standard: the point biserials for all distractors were negative. Further, of the 144 point-biserials for the distractors, all but three were less than -.09. In contrast, 23 items failed to meet the fourth standard that 5% of the examinees select at least one of the item distractors (see Appendix B). As shown in Table 3, the four items which failed to meet either the first or second standards are included within this set of 23 items.

Item Response Item Analyses Results

In the literature review, it was pointed out that the use of item response models rests on three basic assumptions. These assumptions include unidimensionality, local independence of items, and being able to represent mathematically the responses of examinees. In addition to these three general assumptions, additional assumptions specific to a particular model are made. For example, the one-parameter model assumes that the items are equally discriminating and that the influence of guessing is minimal. The two-parameter model assumes that there is essentially no guessing.

Assessing the Assumptions of Item Response Theory

Unidimensionality. The variance accounted for by the first five components yielded by a principal component analysis, the scree test, and Stout's T statistic were used to assess the tenability of the assumption of unidimensionality for the ENGI subtest. The eigenvalues and the corresponding percentages of variance accounted for by the first five components extracted from the correlation matrix containing the phi coefficients and the correlation matrix containing tetrachoric coefficients are reported in Table 4.

Table 4

The First 5 Components: ENGI Subetst

	Phi Correlations			Tetrachoric Correlations		
Comp.	λ	% of Var.	% Δ	λ	% of Var.	% Δ
1	6.385	13.3	-	10.361	21.6	-
2	1.433	3.0	10.3	1.615	3.4	9.2
3	1.275	2.7	0.3	1.451	3.0	0.4
4	1.245	2.6	0.1	1.384	2.9	0.1
5	1.207	2.5	0.1	1.353	2.8	0.1

Note: Var. indicates the variance accounted for by the component extracted.
Δ indicates change in the percentage of variances.

For both phi and tetrachoric correlations, the percentages of variance accounted for by the first component, 13.3% and 21.6% respectively, are more than four times as large as the variances accounted for by the second component. Further, changes in the percentages of variances accounted for by successive components beginning with component [illegible] are small in both cases. Thus, while not large in an absolute sense, the magnitude of the first component is large enough to indicate that there is a dominant component underlying the item responses (Huynh & Ferrara, 1994). Consequently, the variances accounted for by the successive components suggest that the assumption of unidimensionality is met.

This conclusion is supported by the shape of the scree plots. Given the agreement between the scree plots when the phi coefficients were analyzed and when the tetrachoric coefficients were analyzed, only one plot, for the phi coefficients, is presented in Figure 6. Clearly, the first component dominates the remaining components.

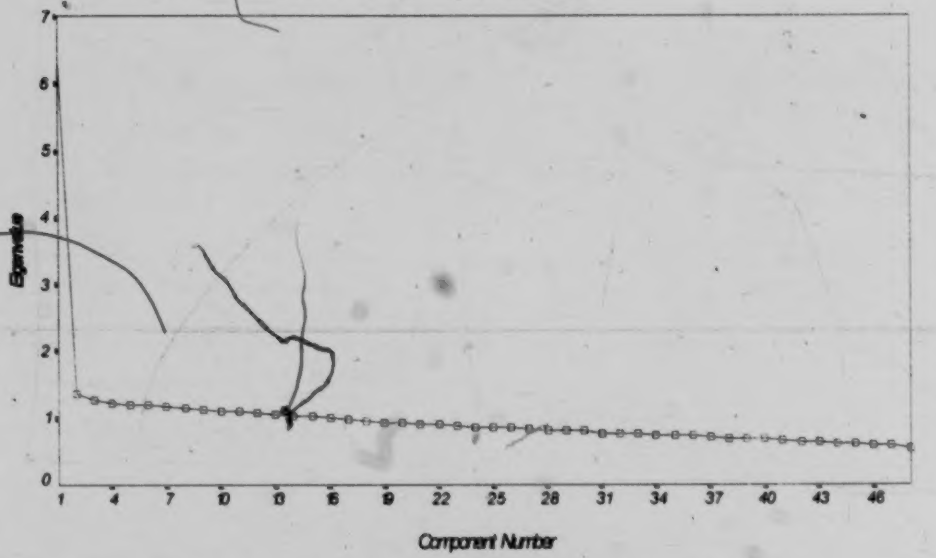


Figure 6: Plot of Eigenvalues Against Component Number: ENGI Subtest

The value of Stout's T statistic, computed using the automatic execution procedure, was 1.023; the associated *p* value was .153. Since this *p* value is greater than .05, the null hypothesis, $d_E = 1$, was not rejected. This implies that the hypothesis of essential unidimensionality can be asserted for ENGI subtest. Thus, the three methods used to assess the assumption of unidimensionality converged to the conclusion that one dominant factor underlies this set of data. Consequently, the assumption of local independence was met as well (Hambleton et al., 1991).

Equal item discrimination. As shown in column 4 of Table 3, the values of the uncorrected point-biserial correlations ranged from .18 to .49. Due to this large variation (Hambleton & Murray, 1983), it was concluded that the assumption of equal discrimination was not met by the 48 items in ENGI subtest. This finding suggests that the one-parameter item response model, which is based upon the assumption of equal discrimination, may not be appropriate for this subtest.

Nonspeededness. To determine the presence of "speededness", the percentage of examinees who completed the last three items in ENGI subtest was examined. Of the 1,232 examinees in the sample, only one did not answer two of the last three items. Therefore, it was concluded that speededness was not a factor that affected the examinees' performances.

Minimal guessing. The performance of low-scoring examinees was examined on the most difficult items. The expectation was that low-scoring examinees would have close to zero performance on the most difficult items if the assumption of guessing is viable. For the ENGI subtest, examinees whose number right scores were below 16 were considered as low-scoring examinees. Sixteen examinees were identified in this category. Their performance on the most difficult items, items 25 (p-value = .40), 26 (p-value = .50), and 27 (p-value = .46), were examined. It was found that 10 of the 16 (62.5%) examinees had a score of 0 for these three items while 5 (31.3%) had a score of 1 and 1 (6.3%) had a score of 2. Therefore, it was concluded that the influence of guessing was minimal because a greater number of examinees than what would have been expected by random guessing were not able to answer these difficulty items. Additional evidence supporting this conjecture can be seen in Figure 5: only 2 out of 1,232 examinees scored less than twelve, the chance level score.

Examination of the Fit of Items to Item Response Models

As with the conventional item analysis, item response item analyses also provide item parameters for the correct or key option for each item. But, unlike the conventional item analysis, no information is provided for the incorrect options. For the one-parameter model, only the difficulty parameter values are provided whereas for the two- and three-parameter models both the difficulty and discrimination parameter values are provided. The three-parameter model also provides an additional parameter, the psuedo guessing parameter, which indicates the probability that an examinee with low ability will answer the item correctly simply by guessing. Estimates of these item parameters are presented in columns 6 to 11 of Table 3.

Given that the ENGI subtest contained 48 items, the chi-square statistic was used to assess the degree of item fit for each of the three item response models. For each model, a significant chi-square ($p \leq .05$) indicates a lack of item fit, that is, the item does not fit the corresponding item response model.

The values of item chi-square statistic for the one-, two-, and three-parameter models revealed that as the number of parameters in the model increased, the number of misfitting items decreased (see Table 3). In the case of one-parameter model, 21 items (3, 6, 9, 11, 13, 24, 25, 27, 28, 30, 34, 39, 41, 42, 44, 45, 50, 55, 58, 64, and 67) were identified as misfitting, while for the two and three-parameter models the numbers were respectively 6 (13, 25, 27, 44, 45, and 55) and 1 (25).

The larger number of misfitting items identified by the one-parameter model appears to be accounted for by the variation in item discrimination. Further examination of point-biserials revealed that the mean item discrimination for the misfitting items was greater than the mean item discrimination for the fitting items (.39 vs .33). In addition, of the 15 items with discrimination indices greater than .39, all but one were misfitting.

The observation that only six of the 48 items were identified by the two-parameter model agrees with the earlier finding that the influence of guessing was minimal. Nevertheless, the results for the three-parameter model, which incorporates a guessing parameter, were better. Further examination of misfitting items revealed no other systematic patterns.

Comparison of Misfitting Items Identified by Conventional and Item Response Item Analyses

The numbers of items identified as misfitting by two or more of the item analyses procedures used are reported in Table 5. In the case of the conventional item analysis, only the two criteria adopted to evaluate the correct option were considered.

First, it should be noted that, with the exception of the one-parameter model, only a small number of items was identified as misfitting. This is likely due, in part, to the care and attention given to the development of these items by Alberta Education.

Turning to misfitting items, only 1 of the 4 items identified in the conventional item analysis failed to meet more than one standard. Three of these items were also identified as misfitting by the one-parameter model, but not by the two- and three-parameter models. Within the item response models, of the 21 items identified as misfitting using the one-parameter model, 6 were identified by the two-parameter model.

And of these six, one item (item 25) was identified by the three-parameter model. This pattern is consistent with the findings reported by Gierl and Hanson (1995). As suggested by Lord (1980) and Traub (1983), the three-parameter model is the most appropriate model for multiple-choice test items. Thus it would appear that the use of one-parameter model may be inappropriate for this set of items, particularly given the lack of homogeneity of item discrimination.

Table 5

Comparison of Misfitting Items for Conventional and Item Response Item Analyses: ENGI Subtest

Model	No. of Items	p-value	Conv. cpbs	pbs	one-	IR two-	three-
Conv.							
p-value	1	-	0	0	1	0	0
cpbis	3		-	1	1	0	0
rpbis	1			-	1	0	0
IR							
one-	21				-	6	1
two-	6					-	1
three-	1						-

Comparison of the Ability Estimates

Results of the comparisons of ability estimates for the ENGI subtest are reported in Table 6. These estimates were computed using the number right scoring algorithm and the scoring algorithms associated with each of the item response models and the finite state scoring model. The correlations among these ability estimates, transformed to a common scale with a mean of 50 and standard deviation of 10, are reported in the upper triangle while the mean absolute differences among each pair of the estimates are reported in the lower triangle.

Table 6

Correlations and Mean Absolute Differences: ENGI Subtest

	NR	θ_1	θ_2	θ_3	λ
NR	-	.991	.984	.986	.957
θ_1	1.10 (.85)	-	.995	.991	.984
θ_2	1.39 (1.09)	.93 (.70)	-	.996	.981
θ_3	1.29 (1.05)	1.07 (.93)	.65 (.66)	-	.970
λ	2.22 (2.02)	1.36 (1.40)	1.36 (1.46)	1.78 (1.77)	-

Note: NR is the number right score, θ_i indicates ability estimates from item response theory, i indicates the number of parameters in the model, and λ is the ability estimate from finite state score theory. Numbers in the brackets indicate the standard deviations.

The correlations among the ability estimates computed using the number right, one-, two-, and three-parameter and finite state score models are quite high with two exceptions. All exceed .98 except for the correlations between the number right and finite state scores, and the three-parameter ability estimates and the finite state scores which were respectively .957 and .970.

Given the high correlations among the estimates, the next aspect examined was the absolute agreement among the scores. This was done by computing the mean absolute difference between pairs of ability estimates which were first transformed to the T-score metric. As shown in the lower triangle of Table 6, the mean absolute differences among pairs of transformed scores ranged from .65 to 2.22. The closer agreements were between the two- and three-parameter item response models (.65) and the one- and two-parameter item response models (.93). The transformed finite state scores agreed most closely with the scores yielded by the one- and two-parameter item response models (1.36 and 1.36 respectively), and less closely with the three-parameter item response model (1.78) and the number right (2.22).

The standard deviations of the absolute deviation scores ranged from .66 to 2.02. With the exception of the pairs involving the finite state scores and the one-, and two-parameter ability estimates, the standard deviations covaried with the corresponding mean absolute deviations. As observed in the case of the mean absolute deviations, the larger variations were between the finite state scores and the three-parameter ability estimates (1.77) and the number right scores (2.02).

The number of misfitting items identified by the one-parameter item response item analysis was far greater than the number of items identified in the conventional item analysis and the two- and three-parameter item response item analyses. However, this did not seem to influence adversely the correlations and the mean absolute deviations between the one-parameter ability estimates with the estimates from other models. In contrast, although the conventional item analysis approach identified a much smaller number of misfitting items than the one-parameter item response item analysis, the mean absolute differences among the pairs involving the number right scores were not smaller than those involving the one-parameter model. Consequently, the failure of the ability estimates to agree closely with one another could not be simply attributed to the presence of misfitting items.

Best Answer English 30 Items

The results presented in this section are based on the analyses of the 20 items in the English 30 examination which required examinees to select the best answer from among the four options. Since the options need to be ordered based upon the degree of correctness, they are not independent of one another. Consequently, the assumption of option independence is not met. For convenience, this subtest is referred to as ENGDB where DB denotes dependence among options due to best answer.

Confirmation of Option Order

Prior to conducting the main analyses, an analysis was conducted to verify if the items could be ordered in terms of the degree of correctness. Two approaches were used to verify this. The first involved judgements by a panel of teachers who were teaching English 30. The second involved an examination of option p-values for each option as done in empirical weighting (Smith, 1985, p. 218). Teachers' ratings were then compared with the empirical weightings.

Seven teachers were approached. The teachers were selected and approached by the Curriculum Coordinator for English in a large, metro school district in Alberta. The selection criteria were: (a) teaching experience of at least 2 years and (b) presently teaching English 30.

All seven teachers agreed to participate. Each was presented a package which contained (a) a letter explaining the purpose of obtaining teachers' ranking of the items, (b) English 30 Examination Readings Booklet, (c) examination items, (c) instructions for ranking and (d) ranking sheet. The teachers were asked to work independently, and to return all materials to the Curriculum Coordinator for English.

Five teachers completed the task. Review of the ratings of the five teachers (see Appendix C) revealed that one judge disagreed with the other four judges on 7 of 20 items (35%) in terms of the best answer. For example, for item 8, this judge ranked option A as the least best option while the other four judges ranked it as the best option. Thus, this judge was removed, leaving a panel of 4 judges. The mean ratings of the four judges for each option within each item are reported in Table 7.

Table 7

Judges Ratings Versus Empirical Weighting: ENGDB Subtest

	Option					Option			
Item	A	B	C	D	Item	A	B	C	D
5.	1.00	2.00	3.50	3.50[a]	8.	1.00	2.75	3.75	2.50
	1.0	1.0	4.0	3.0[b]		1.0	2.0	4.0	3.0
12.	2.50	3.75	1.00	2.75	14.	1.00	2.25	2.75	4.00
	3.0	4.0	1.0	2.0		1.0	2.0	3.5	3.5
16.	3.50	1.00	3.00	2.50	18.	3.50	3.00	1.00	2.50
	4.0	1.0	3.0	2.0		3.5	3.5	1.0	2.0
21.	2.25	3.25	3.50	1.00	29.	3.00	3.50	2.50	1.00
	2.0	3.0	4.0	1.0		4.0	3.0	2.0	1.0
31.	1.00	3.75	3.25	2.00	32.	3.25	3.75	1.00	2.00
	1.0	4.0	3.0	2.0		4.0	3.0	1.0	2.0
36.	2.00	1.00	3.00	4.00	46.	3.25	2.25	3.25	1.00
	2.0	1.0	3.0	4.0		3.0	4.0	2.0	1.0
47.	2.00	3.40	2.60	1.00	51.	2.50	3.00	1.00	3.50
	2.0	4.0	3.0	1.0		4.0	2.0	1.0	3.0
52.	2.75	3.00	1.00	3.25	53.	3.75	3.25	1.00	2.00
	3.0	3.0	1.0	3.0		4.0	3.0	1.0	2.0
54.	2.00	1.00	3.67	3.33	60.	3.33	2.33	1.00	3.00
	4.0	1.0	2.0	3.0		3.0	2.0	1.0	4.0
62.	2.75	1.00	3.00	3.25	70.	1.00	3.75	2.40	2.80
	3.0	1.0	2.0	4.0		1.0	4.0	3.0	2.0

[a] Indicates the mean option rank for the four judges.

[b] Indicates the rankings using p-values.

A low mean indicates a more correct option while a high mean indicates a less correct option. For some items, one judge ranked only one or two options. Consequently, for these items the means are not a multiple of .25. The number below each mean is the rank assigned to each option using the empirical p-values computed from the sample of 1,232 examinees. For example, for item 5, the p-values for options A and B were equal and the highest. They were followed, in turn, by the p-values for option D and then option C. Hence the ranking of 1, 1, 4, and 3.

While the number of judges is small, the agreement among the four judges was perfect for the first or best options ($\bar{X} = 1.00$) and nearly perfect for the second or second best options as reflected by mean ratings. There was less agreement among the judges for the remaining two options for several items. For example, for item 5 all the judges agreed on the best (option A) and second best (option B) options; they were evenly split regarding the degree of correctness of the remaining two options.

Examination of the empirical weights defined in terms of the p-values revealed three items (5, 14, and 18) with tied ranks. Further, inspection of the p-values suggested that, while different, several of the p-values for the third and fourth best options were quite close.

Taken together, the judges and empirical rankings were the same for 8 out of the 20 items. If just the first two options are considered, there was an agreement for 12 items. In light of these results, the application of the partial credit model was questionable. However, the decision was taken to conduct this analysis to examine if the partial credit analysis was sensitive to the less than perfect ordering of options. The results were compared to those obtained using the number right, one-, two-, and three-parameter item response, and finite state scoring models.

Subtest Characteristics

The distribution of number right scores for the ENGDB subtest is presented in Figure 7. The mean for this set of items was 13.09 (66.0%) and the standard deviation was 3.42 (16.9 %). The Cronbach's alpha was .67 and the standard error of measurement was 1.96 (9.3%). Given that the number of items in this subset is less than half of the number of items in the subtest of 48 items with independent options, the value for Cronbach's alpha indicates a high degree of internal consistency or item homogeneity. Applying the Spearman-Brown Prophecy Formula, the expected internal consistency of a 48 item test obtained by adding 28 parallel items to the initial 20 items is .83. The skewness coefficient was -.21 and the kurtosis coefficient was -.60. Again, these values indicate that the distribution was somewhat skewed to the left and was flatter than the normal distribution.

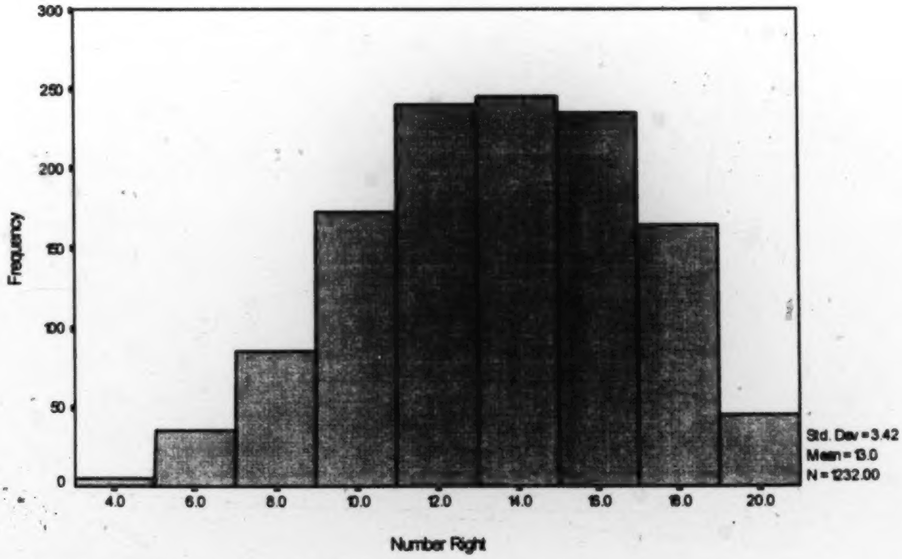


Figure 7: The Distribution of Number Right Scores: ENGDB Subtest

Item Characteristics

The estimated item parameters obtained from the conventional, item response, and partial credit item analyses for the ENGDB subtest are presented in Table 8. In addition, the incorrect options with p-values less than .05 are listed. The difficulty parameters for the one-, two-, and three-parameter models are in normal unit deviates while the difficulty parameters for the partial credit model are in logits. For each model, misfitting items are identified by an asterisk (*).

Table 8:

Conventional, Item Response and Partial Credit Item Analyses Results: ENGDB Subtest

	Conventional				IRT Parameter Estimates							
					1PL	2PL		3PL			PC	
item	p	cpbs	pbs	option	b	b	a	b	a	c	b^e	b^r
5	0.67	0.28	0.40	c	-1.11*	-0.96	0.51	-0.34	0.61	0.23	-0.12	-0.16
8	0.57	0.32	0.45	c	-0.42*	-0.35	0.58	0.14	0.74	0.19	0.00	-0.07
12	0.51	0.27	0.40	✓	-0.06*	-0.08	0.46	0.58	0.62	0.22	0.31	0.24
14	0.66	0.22	0.35	✓	-1.03	-1.10	0.39	0.02	0.55	0.32	0.05	-0.02
16	0.75	0.26	0.37	✓	-1.75*	-1.47	0.52	-0.78*	0.63	0.27	-0.12	-0.17
18	0.51	0.23	0.36	✓	-0.03	-0.05	0.38	0.71	0.53	0.22	0.67	0.39*
21	0.70	0.26	0.38	✓	-1.32*	-1.20	0.47	-0.38	0.61	0.28	0.10	0.04
29	0.69	0.19*	0.32	✓	-1.23	-1.39	0.36	-0.30	0.47	0.30	0.07	0.12
31	0.80	0.20	0.31	c	-2.15*	-2.08*	0.43	-1.10	0.53	0.33	-0.59	-0.64
32	0.71	0.23	0.35	✓	-1.42	-1.40	0.43	-0.44	0.55	0.30	-0.05	0.03
36	0.57	0.29	0.42	✓	-0.45*	-0.40	0.51	0.30	0.72	0.25	0.04	-0.02
46	0.73	0.33	0.44	✓	-1.58*	-1.11*	0.67	-0.62*	0.81	0.23	0.08	0.13
47	0.76	0.29	0.40	b	-1.76*	-1.32*	0.61	-0.58*	0.80	0.31	-0.15	-0.21
51	0.46	0.23	0.37	✓	0.28	0.27	0.38	0.94*	0.51	0.19	0.58*	0.54*
52	0.89*	0.21	0.30	a,b,c	-3.22*	-2.43	0.59	-1.89	0.66	0.26	-1.17	-0.37
53	0.44	0.32	0.45	✓	0.43*	0.31*	0.58	0.72*	0.83	0.17	0.33	0.26
54	0.57	0.20	0.33	✓	-0.41*	-0.50*	0.35	0.92	1.02	0.41	0.37	0.39*
60	0.59	0.19*	0.33	✓	-0.53	-0.67	0.33	0.68	0.52	0.33	0.14	0.17
62	0.72	0.15*	0.28	d	-1.46	-1.98	0.29	-0.60	0.37	0.32	-0.45	-0.50
70	0.68	0.23	0.36	b	-1.20	-1.20	0.42	-0.38	0.50	0.26	-0.10	-0.16

b^e Ability estimates computed when the empirical weightings were used.
b^r Ability estimates computed when the teachers' ratings were used.

Conventional Item Analysis Results

The p-values for the correct options, shown in the second column of Table 8, ranged from .44 to .89. All items met the minimum standard of .30. One item, item 52 (p-value = .89), exceeded the maximum standard of .85. The corrected point-biserial item discrimination indices for the correct (best) option ranged from .15 to .35. Three items, 29 (${}_{c}r_{pbs} = .19$), 60 (${}_{c}r_{pbs} = .19$), and 62 (${}_{c}r_{pbs} = .15$), failed to meet the minimum standard of .20. Thus, at this point, four items, 29, 52, 60, and 62, failed to meet at least one of the two standards set for the correct option. Using the uncorrected point-biserial coefficient, all the items met the minimum standard of .20 for the discrimination index.

All 20 items met the third standard: all the point-biserials for distractors were negative (see Appendix D). Further, of the 60 point-biserials for the distractors, all but two were less than -.09. In contrast, 7 items failed to meet the fourth standard that at least five percent of the students select at least one of the item's distractors. When all four standards are used, 9 items failed to meet at least one of the standards set for either the correct option or the distractors.

Item Response Item Analysis Results

Assessing the Assumptions of Item Response Theory

Unidimensionality. The eigenvalues and the corresponding percentages of variance accounted for by the first five components extracted from the correlation matrix containing phi coefficients and the correlation matrix containing tetrachoric coefficients are reported in Table 9.

Table 9

The First 5 Components: ENGDB Subtest

	Phi Correlations			Tetrachoric Correlations		
Comp.	λ	% of Var.	% Δ	λ	% of Var.	% Δ
1	3.092	14.1	-	4.069	20.3	-
2	1.121	5.1	9.0	1.198	6.0	14.3
3	1.082	4.9	0.2	1.139	5.7	0.3
4	1.073	4.8	0.1	1.090	5.2	0.5
5	1.013	4.6	0.2	1.008	5.0	0.2

Note: Var. indicates the variance accounted for by the component extracted.
Δ indicates change in the percentage of variances.

As shown, the percentage of variances accounted for by the first component (14.1% and 20.3%) using both phi and tetrachoric correlations are approximately three times larger than the variances accounted for by the second component (5.1% and 6.0%). Further, changes in the percentages of variance accounted for between successive components, beginning with the second component, are small. These findings suggest that the assumption of essential unidimensionality is met. This conclusion is supported by the shapes of the scree plots (see example, Figure 8).

The value of Stout's T statistic computed for this set of data was .685 and the associated *p* value was .247. Together, these values suggest that there is essentially one dimension underlying the examinees responses. Thus, the assumption of unidimensionality and, consequently, the assumption of local independence are met in the set of 20 best answer English items.

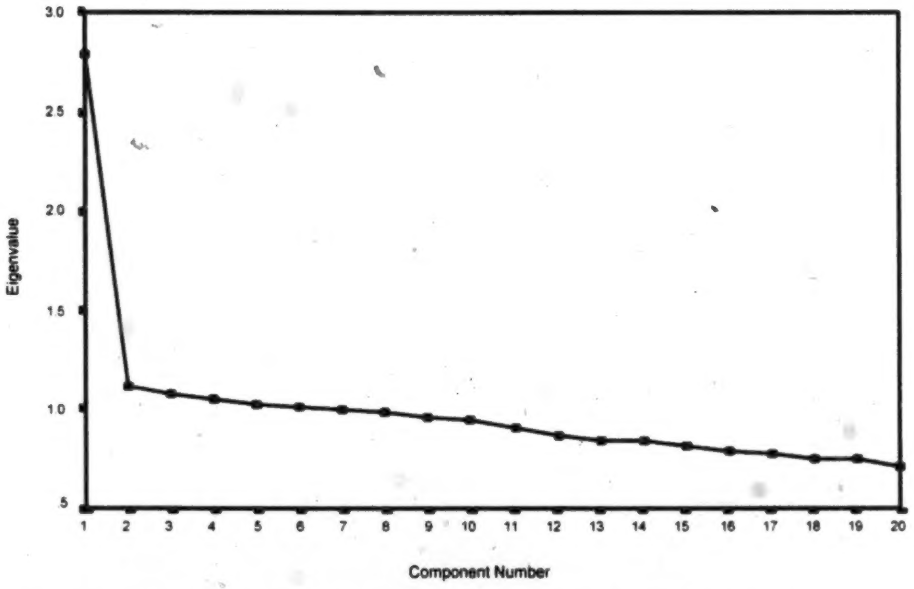


Figure 8: Plot of Eigenvalues Against Component Number: ENGDB Subtest

Equal item discrimination. As shown in the third column of Table 8, the values of the uncorrected point-biserial correlations ranged from .27 to .45. Due to this large variation, it was concluded that the assumption of equal discrimination was not met by the 20 items in the ENGDB subtest. This suggests that the application of one-parameter item response model and partial credit model, which are based upon the assumption of equal item discrimination, may not be appropriate for this subtest.

Nonspeededness. The assumption of nonspeededness was assessed by examining the percentage of examinees who completed the test. Of the 1,232 examinees in the sample, only 2 (0.2%) did not answer one of the last three items in the ENGDB subtest.

Thus, it was concluded that speededness was not a factor that affected the examinees' performances.

Minimal guessing. Forty-two examinees whose number right scores were less than 7 were identified as low-scoring examinees for this subtest. It was expected that these examinees would have a close to zero performance on the most difficult items. Their performance on the three most difficult items, items 18 (p-value = .51), 51 (p-value = .46), and 53 (p-value = .44), were examined. It was found that 31 of these 42 examinees (73.8%) had a score of 0, while 8 (19.1%) had a score of 1, and 3 (7.1%) had a score of 2. Accordingly, it was concluded that the influence of guessing was minimal. Additional evidence supporting this finding can be seen in Figure 7; only 17 examinees scored at or below the chance score of five.

Examination of the Fit of Items to Item Response Models

The values of the item parameters for each of the one-, two-, and three-parameter item response models are reported in columns 6 to 11 of Table 7. Using the chi-square statistic as a test of fit, 12 items (5, 8, 12, 16, 21, 31, 36, 46, 47, 52, 53, and 54) were identified as misfitting the one-parameter model, while 5 items (31, 46, 47, 53, and 54) and 5 items (16, 46, 47, 51, and 53) were identified as misfitting, respectively, the two- and three-parameter item response models ($p \leq .05$).

Like in the case of ENGI subtest, the larger number of misfitting items identified by the one-parameter model appears to be accounted for by the variations in item discrimination. Examination of point-biserial correlations revealed that the mean item discrimination for the misfitting items was greater than the mean item discrimination for the fitting items (.39 vs .34). Furthermore, all 7 items with discrimination index greater than .39 were misfitting. Similarly, the observation that the same number of items were identified as misfitting the two- and three-parameter models concurs with the initial finding that the influence of guessing was minimal.

Partial Credit Item Analysis Results

As mentioned previously, the items included in the ENGDB called for the best answer. Consequently, the partial credit method of scoring was considered. Since this model is an extension of the one-parameter model, it is also based on the assumption that only the difficulty parameter plays role in determining the examinees' responses to the test items. Thus, like the one-parameter model, only the difficulty parameter is provided. The parameter estimation and, subsequently, ability estimation were done twice: first, using empirical weightings and second, using teachers' ratings. The values of the difficulty parameters are presented in the last two columns of Table 13.

Examination of the Fit of Items to the Partial Credit Model

The fit of items to the partial credit model was determined using the infit and outfit statistics. An item with either an infit or outfit statistic greater than 3 is considered as misfitting the partial credit model. The results of the item analyses using empirical weightings revealed that one item, 51, was identified as misfitting the partial credit model whereas using teachers' ratings, three items, 18, 51, and 54, were identified as misfitting.

While item 51 is one among the most difficult items, items 18 and 54 had difficulty and discrimination parameters which were within the average range. Consequently, the possible cause of the item-misfit could not be clearly established.

Comparison of Misfitting Items Identified by the Conventional, Item Response, and Partial Credit Item Analyses

The numbers of items identified as misfitting by two or more of the item analyses procedures are reported in Table 10.

Table 10

Comparison of Misfitting Items for the Conventional, Item Response, and Partial Credit Item Analyses: ENGDB Subtest

			Conv.			IR		PC	
Model	Items	p-value	cpbis	pbis	one-	two-	three-	Emp.	Trat.
Conv.									
p-value	1	-	0	0	1	0	0	0	0
cpbis	3		-	0	0	0	0	0	0
pbis	0			-	0	0	0	0	0
IR									
one-	12				-	5	4	0	1
two-	5					-	3	0	1
three-	5						-	1	0
PC									
Emp.	1							-	1
Trat.	3								-

Note: Emp indicates the empirical weighting using the p-values.
Trat indicates the teachers' rankings.

Four items failed to meet one of the conventional item analysis standards for the correct option. One of these items was also identified by the one-parameter model. Within the item response models, of the 12 items identified as misfitting using the one-parameter model, 5 were identified by the two-parameter model, and 4 by the three-parameter model. Three of the five items identified as misfitting the two-parameter model also misfit the three-parameter model. Thus, like the case for the ENGI subtest, there was closer agreement between the two- and three-parameter models than between each of these models and the one-parameter model.

Comparison of the Ability Estimates

The results of the comparisons of ability estimates for the ENGDB subtest are reported in Table 11. These estimates were computed using the number right scoring algorithm and the scoring algorithms associated with each of the one-, two-, and three-parameter models, partial credit model, and the finite state scoring model. In the case of the partial credit model, no ability estimates were provided for the 12 examinees with perfect scores although all 1,232 examinees were included in the estimation of item parameters and ability estimates. Consequently, these examinees were also eliminated from other analyses in order to facilitate the comparison among the different scoring approaches using the same number of examinees in each case. Given the similarity between the ability estimates obtained when the teachers' ratings were used and when the empirical weightings were used, only the comparisons involving the empirical weightings are reported in Table 11.

Table 11

Correlations and Mean Absolute Differences Among Ability Estimates: ENGDB Subtest

	NR	θ_1	θ_2	θ_3	λ	P_c
NR	-	.999	.991	.985	.955	.902
θ_1	.41 (.29)	-	.992	.988	.963	.915
θ_2	1.09 (.80)	1.01 (.75)	-	.997	.957	.908
θ_3	1.33 (1.02)	1.24 (.94)	.56 (.53)	-	.952	.909
λ	2.15 (1.99)	1.81 (1.96)	2.01 (1.98)	2.10 (2.13)	-	.911
P_c	3.29 (2.91)	3.10 (2.70)	3.17 (2.81)	3.17 (2.78)	2.80 (3.04)	-

Note: NR is the number right score, θ_i indicates ability estimates from item response theory, i indicates the number of parameters in the model, P_c indicates estimates based on the partial credit model, and λ is the ability estimate from finite state score theory.

As shown in the upper triangle of Table 11, the correlations among the ability estimates computed using the number right and one-, two-, and three-parameter scoring algorithms and transformed to a common scale ($\mu = 50$, $\sigma = 10$) exceed .98. In contrast, while still high, the correlations between the finite state score estimates with these estimates are, somewhat, lower (.955, .963, .957, and .952). Lastly the correlations between the ability estimates computed using the partial credit scoring algorithm and each of the other estimates are systematically lower than the correlations among the other ability estimates.

The pattern of mean absolute differences among the pairs of transformed scores corresponds to the pattern observed for correlations. The mean absolute differences computed for the number right score, one-, two-, and three-parameter ability estimates ranged from .41 to 1.33 while the mean absolute differences between the finite state ability estimates and these estimates ranged from 1.81 to 2.15. The range of mean absolute differences for the partial credit ability estimates and the other estimates was 2.80 to 3.29.

The pattern among the variabilities of the absolute deviations followed the same pattern as the means. The standard deviations of the absolute differences between the number right estimates and the one-, two-, and three-parameter estimates ranged from .29 to 1.02, while the standard deviations of the absolute differences between the finite state score estimates ranged from 1.96 to 2.13. The corresponding range for the partial credit model was 2.70 to 3.04

As observed in the case of the ENGI subtest, a larger number of misfitting items was identified for the one-parameter model than for the other models. Again, the correlations and mean absolute differences based on ability estimates obtained from the one-parameter item response model suggested that the ability estimates were not influenced by this lack of fit. Therefore, it appears that lack of fit of the items had little or no influence on the rankings and absolute deviations.

Discussion

Comparison of the correlations among the ability estimates yielded by the number right and one-, two-, and three-parameter item response scoring algorithms for the ENGDB subtest with the corresponding correlations for the ENGI subtest reveals that they are essentially the same (cf. Tables 6 and 11). In contrast, with the exception of the correlations between the number right and finite state scores, which remained the same, the correlations of the finite state scores with these estimates for the ENGDB subtest are somewhat lower than for the ENGI subtest (.957 vs .952; .984 vs .963; .981 vs .957 and .970 vs .952).

The pattern of the differences between the mean of the absolute differences for the ENGDB subtest and corresponding means for the ENGI subtest is consistent with that for the correlations (cf. Tables 6 and 11). The two sets of mean absolute deviations computed from the number right and the one-, two-, and three-parameter estimates are, with the exception of the pair for the number right and one-parameter ability estimates, approximately equal. The mean absolute deviations between the number right and the finite state score ability estimates are approximately the same; the remaining three mean absolute differences, involving the finite score ability estimates, are somewhat larger for the ENGDB subtest than for the ENGI subtest (2.02 vs 1.99; 1.46 vs 1.98; and 1.77 vs 2.13, respectively). Lastly, the pattern for the standard deviations of the absolute differences is similar to that observed for the mean absolute deviations, both within the ENGDB subtest and across the ENGDB and ENGI subtests.

Taken together, the correlations among the ability estimates and means and standard deviations of the absolute difference between ability estimates indicate that the agreement among the number right scores and the ability estimates yielded by the one-, two-, and three-parameter item response models is good. While not large, the differences between these estimates and the finite state score estimates suggest that the finite state scoring algorithm was sensitive to the violation of the assumption independent options brought about by asking for the best instead of the correct answer. The correlations were slightly lower and the means and standard deviations of the absolute deviations were slightly higher. However, the difference was not as high as expected. This failure is likely due to the observation that while the best two options were rankable, the others were not. Indeed, this may well be the explanation for the poor performance of the partial credit model; the four options were not clearly ranked.

The presence of items which failed to meet the standards in a conventional item analysis or which were identified by one or more of the item response models as misfitting did not appear to influence the differences observed between the ability estimates. First, with the exception of the one-parameter model, only a small number of misfitting items were identified by each procedure. Further, the results of the comparisons in which the one-parameter ability estimates were involved are not unlike the results for the remaining comparisons. Consequently, item "misbehavior" did not appear to be an influential factor.

A rival explanation for the observed differences between the correlations, means, and standard deviations for the ENGI subtest and for ENGDB subtest is the difference in the number of items in each subtest. This difference would affect the reliabilities of the two subtests, with the shorter having the lower reliability. However, it would be expected that all correlations, means, and standard deviations would be affected. Clearly, the values of these statistics were not influenced for the pairs formed from the number right and one-, two-, and three-parameter estimates. Thus, it was concluded that the lower correlations and higher means and standard deviations for the pairs in which the finite state scores were a member were not due to attenuation brought about by lower reliability.

CHAPTER 5: RESULTS BASED ON THE TEST OF TEST-WISENESS ITEMS

Results based on the analyses of the Test of Test-Wiseness items are presented in this chapter. As in the case for the English 30 Examination, two subtests were formed. The first subtest contained 32 items which satisfied the assumption of option independence. The second subtest contained 18 items for which there was either a pair of opposite options or a pair of similar options, thereby violating the assumption of option independence.

Like the previous chapter, the present chapter is organized into three main sections. The first two sections contain results for the two subtests formed from the Test of Test-Wiseness. The last section contains a discussion of the findings for the two subtests, with reference to the findings for the two subtests presented in the previous chapter.

Test-Wiseness Items Which Satisfied the Assumption of Option Independence

The results presented in this section are for the 32 items in the Test of Test-Wiseness which satisfied the assumption of independence among the options. These items required the examinees to select the correct options. For convenience, this subtest is referred to as TTWI, where I denotes independent options.

Subtest Characteristics

The distribution of number right scores for the TTWI subtest is presented in Figure 9. The mean score for this set of items was 17.69 (53.3%) and the corresponding standard deviation was 3.99 (12.5 %). The Cronbach's alpha was .59 and the standard error of measurement was 2.57 (8.0%). Given the number of items in the test, the internal consistency of items was not high. The expected internal consistency of 48 item test obtained by adding 16 parallel items is .68. This value is still somewhat low. This lower value is likely attributable to the observations that the Test of Test-Wiseness contained items from four subject areas. Consequently, it might be expected that the internal consistency of this set of items would be somewhat lower than for the ENGI subtest which assessed only one subject area. The skewness coefficient was -.09 and the kurtosis coefficient was .01. These values show that the shape of the distribution was essentially normal.

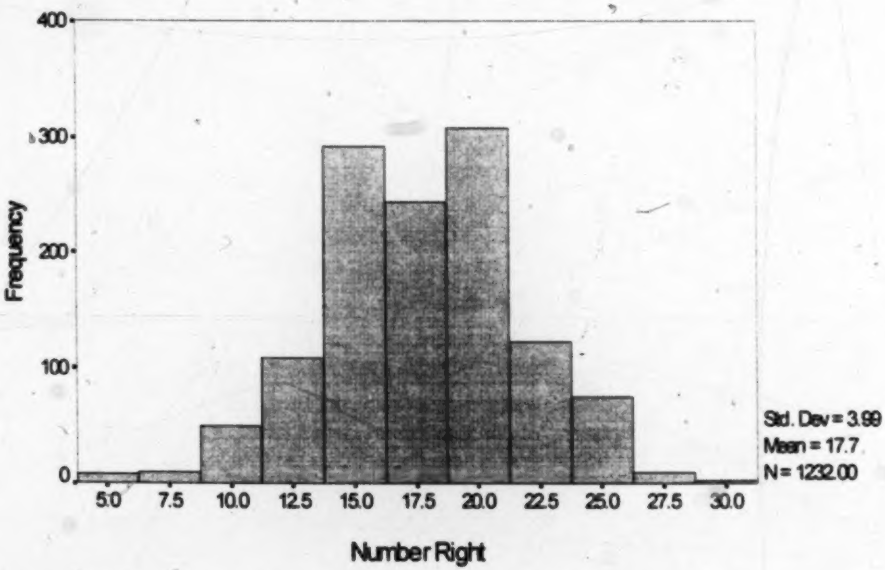


Figure 9: Distribution of Number Right Scores: TTWI Subtest

Item Characteristics

The estimated item parameters obtained from the conventional and item response item analyses are listed in Table 12 for each of the 32 items in the TTWI subtest. The results from the conventional analysis include the p-values and uncorrected point-biserial correlations for the correct options. The corrected point-biserial correlations were not used for this subtest because the developers of this test used the uncorrected coefficients (Rogers & Bateson, 1989; Rogers & Wilson, 1993). The incorrect options for which the p-values failed to meet the minimum level of .05 are also listed. Misfitting items for each of the models are identified by an asterisk (*).

Conventional Item Analysis Results

The results obtained from the conventional item analysis are presented in columns 2 to 4 of Table 12. The item difficulties for the correct options, shown in the second column, ranged from .19 to .92. While item 30 (p-value = .26) and item 39 (p-value = .19) did not meet the minimum standard of .30, item 2 (p-value = .92), item 11 (p-value = .88), and item 13 (p-value = .90) exceeded the maximum standard of .85. Thus, five items failed to meet the first standard set by Alberta Education for the item difficulty.

The uncorrected point-biserial coefficients ranged from .15 to .44 (see Column 4, Table 12). Two items (item 3, $r_{bis} = .15$, and item 30, $r_{bis} = .19$) failed to meet the minimum value of .20. Thus, at this point, six different items failed to meet at least one of the two standards set for the correct option.

Table 12

Conventional and Item Response Item Analysis Results: TTWI Subtest

Item	Conventional			IRT Parameter Estimates					
				1PL	2PL		3PL		
	pvalue	pbs	option	b	b	a	b	a	c
2	0.92*	0.24	a,b,c	-5.47*	-3.15	0.51	-2.58*	0.56	0.24
3	0.70	0.15*	b	-1.96	-3.25	0.16	-0.56	0.19	0.34
5	0.53	0.24	✓	-0.25	-0.29	0.24	1.52	0.41	0.34
7	0.68	0.22	c	-1.70	-2.09	0.22	-0.39	0.26	0.30
8	0.56	0.22	c	-0.58	-0.73*	0.22	1.07*	0.27	0.29
9	0.44	0.24	✓	0.55	0.63	0.23	2.07	0.66	0.36
11	0.88*	0.25	b,c	-4.41*	-2.60	0.50	-1.91*	0.58	0.26
13	0.90*	0.23	a	-4.92*	-2.95	0.48	-2.28	0.56	0.25
14	0.39	0.25	✓	1.00*	1.04	0.26	2.08*	0.56	0.28
17	0.48	0.26	a	0.16	0.16	0.25	1.61	0.93	0.40
19	0.55	0.22	✓	-0.43	-0.56	0.21	1.85	0.67	0.47
22	0.49	0.28	✓	0.11*	0.10	0.29	1.47	0.74	0.37
24	0.68	0.29	✓	-1.69	-1.38*	0.34	-0.27*	0.43	0.29
25	0.72	0.22	✓	-2.10	-2.62	0.21	-1.08	0.22	0.29
28	0.33	0.34	✓	1.60*	1.00	0.46	1.33	0.79	0.16
29	0.69	0.37	c	-1.85*	-1.04	0.54	-0.51*	0.65	0.21
30	0.26*	0.19*	✓	2.38*	2.73*	0.23	2.01	1.47	0.22
31	0.54	0.24	✓	-0.33	-0.37	0.24	1.53	0.38	0.35
33	0.57	0.39	a	-0.65*	-0.37	0.54	0.36*	0.83	0.27
35	0.59	0.32	✓	-0.86*	-0.64	0.37	0.70	0.72	0.38
36	0.31	0.23	✓	1.80	1.81	0.27	2.33	0.55	0.21
39	0.19*	0.25	✓	3.24*	2.42	0.38	1.92	1.20	0.13
40	0.65	0.34	✓	-1.40*	-0.91	0.45	-0.03	0.62	0.29

Table 12 continued

	Conventional			IRT Parameter Estimates					
				1PL	2PL		3PL		
Item	p-value	pbs	option	b	b	a	b	a	c
41	0.55	0.21	✓	-0.43	-0.60	0.19	1.85	0.30	0.36
42	0.44	0.44	✓	0.56*	0.26*	0.67	0.61*	0.85	0.14
44	0.56	0.35	✓	-0.57*	-0.37*	0.46	0.49*	0.78	0.29
45	0.34	0.21	✓	1.46	1.69	0.23	2.24	0.71	0.27
46	0.37	0.22	✓	1.23	1.46	0.23	2.36	0.64	0.30
47	0.63	0.33	✓	-1.20*	-0.82	0.43	0.06	0.56	0.27
48	0.66	0.32	✓	-1.50*	-1.07	0.40	-0.12	0.53	0.28
49	0.33	0.24	✓	1.57	1.66	0.25	2.19	0.78	0.28
50	0.35	0.34	a	-2.44*	-1.41*	0.52	-0.61*	0.66	0.30

Three items (30, 36, and 46) failed to meet the third standard: the point-biserial coefficient for at least one distractor was positive (see Appendix E). Further, of the remaining 93 distractors, 5 had point-biserial values between -.01 and -.05 while 8 had point-biserial values between -.06 and -.09 and the remaining distractors had values less than -.09.

Ten items (see column 5 of Table 12) failed to meet the fourth standard that at least 5% of the examinees select each distractor. Four of these items also failed to meet at least one of the two standards for the correct option. Consequently, when all four standards for evaluating correct options and distractors are examined, 14 of the 32 items fail to meet the standards adapted for this study.

Item Response Item Analyses Results

Assessment of the Assumptions of Item Response Theory

Unidimensionality. The variance accounted for by the first five components yielded by a principal component analysis, the scree test, and Stout's T statistic were used to assess the assumption of unidimensionality for the TTWI subtest. The eigenvalues and the corresponding percentages of variance accounted for by the first five components extracted from the correlation matrix containing phi coefficients and the correlation matrix containing tetrachoric coefficients are presented in Table 13.

Table 13

The First 5 Components: TTWI Subtest

	Phi Correlations			Tetrachoric Correlations		
Comp.	λ	% of Var.	% Δ	λ	% of Var.	% Δ
1	2.560	8.0	-	3.590	11.2	-
2	1.439	4.5	3.5	2.228	7.0	4.2
3	1.270	4.0	0.5	1.452	4.5	2.5
4	1.214	3.8	0.7	1.392	4.4	0.1
5	1.192	3.7	0.1	1.330	4.2	0.2

Note: Var. Indicates the variance accounted for by the component extracted.
Δ indicates change in the percentage of variances.

The percentages of variance accounted for by the first component, 8.0% and 11.2% respectively, are small for both the phi and tetrachoric correlations. In both cases, the changes between the percentages of variance accounted for by the first and second components are relatively larger than the changes in the percentages of variances accounted for by successive components beginning with the second component. Huyhn and Fererra (1994) observed that good estimates of item parameters can be obtained when the first component accounts for less than 10% of the total variance, provided that the percentage of variance accounted for by the first component is somewhat larger than the remaining components. The variance accounted for by the first component and the difference between this component and the second component suggests that the assumption of unidimensionality is met. While the scree test suggests the possibility of a second factor, there still appears to be a dominant first component. However, Stout's T statistic revealed that there may be more than one factor. The value of Stout's T statistic was 2.28. Since the associated *p* value is less than .005, the null hypothesis that $d_E = 1$ is not tenable, implying that the data set has more than one dimension.

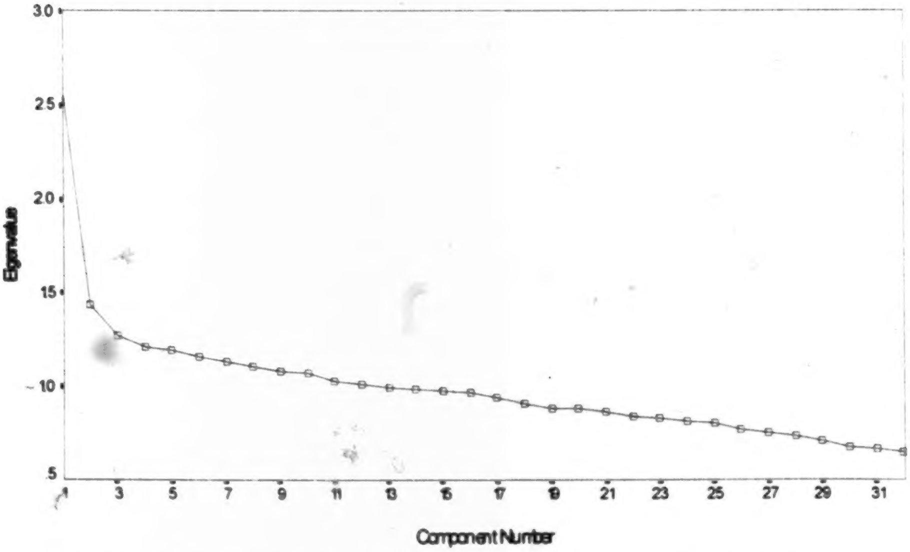


Figure 10: Plot of Eigenvalues Against Component Number: TTWI Subtest

Given the ambiguity in the results of these testing procedures, the decision was taken to proceed with item analyses using the one-, two-, and three-parameter item response models. If the number of misfitting items was proportionally larger than the numbers found in the analyses of ENGI responses reported in Chapter 4, then this finding would be attributable to the lack of unidimensionality for the set of items included in the TTWI subtest.

Equal item discrimination. The values of the uncorrected point-biserial correlations ranged from .15 to .44. Since the difference between the highest and lowest values exceeded .15, it was concluded that the assumption of equal discrimination was not met in the TTWI subtest. This finding suggests that the one-parameter item response model may not be appropriate for the TTWI subtest.

Nonspeededness. The percentage of examinees who completed the test was examined in order to assess the assumption of nonspeeded test administration. Of the 1,232 examinees in the sample, only 6 did not answer one of the last three items in the subtest, and 5 did not answer two of the last three items. Therefore, it was concluded that speededness was not a factor that affected examinees' performances.

Minimal guessing. The performance of low-achieving examinees on the three most difficult items, items 30 (p-value = .26), 36 (p-value = .31), and 39 (p-value = .19), was examined. There were 50 examinees whose number right scores were less than 11. Of these examinees, 36 (72.0%) had a total score of 0 for these items and 14 (28.0%) had a total score of one. It was, therefore, concluded that the influence of guessing was minimal because too many examinees had a score of zero.

Examination of the Fit of Items to the Item Response Models

The fit of items to each of the item response models was assessed using the chi-square statistic because the TTWI subtest contained 32 items. Again, for each model, a significant chi-square indicated a lack of item fit.

The number of items identified as misfitting the one-, two-, and three-parameter models were, respectively, 17 (2, 11, 13, 14, 22, 28, 29, 30, 33, 35, 39, 40, 42, 44, 47, 48, and 50), 6 (8, 24, 30, 42, 44, and 50), and 10 (2, 8, 11, 14, 24, 29, 33, 42, 44, 50). Although the tenability of the assumption of unidimensionality was not clearly established by the three methods used in this study to examine dimensionality, the numbers of misfitting items, which are proportionaly greater than the corresponding numbers for the ENGI subtest, may have influenced the results. With the one-parameter model, the poor fit of the items to the model may also be due to the variation in item discrimination.

The item difficulties and discrimination indices determined in the conventional item analysis were examined in an attempt to determine whether these values would help to explain why some of the items did not fit the item response models. Again, no clear pattern emerged.

Comparison of Misfitting Items Identified by the Conventional and Item Response Item Analyses

The numbers of items identified as misfitting by two or more of the item analyses procedures used are presented in Table 14. In the case of conventional analysis, the two criteria adopted to evaluate the correct option were used while in the case of each of the item response models, the chi-square statistic was used.

Five items (2, 3, 11, 13, and 30) failed to meet one of the standards for conventional item analysis for the correct option. Of these, one item (item 30) failed to meet both standards. Four of the five items were also identified as misfitting by the one-parameter model and three were identified as misfitting by the three-parameter models. Within the one-parameter model, of the 17 items identified as misfitting the one-parameter model, 4 were also identified by the two-parameter model, and 8 by the three-parameter model. Further, of the 6 items identified by the two-parameter model, 5 were also identified by the three-parameter model. Three items (42, 44, and 50) were identified as misfitting each of the item response models. The two-parameter model provided a better fit of the items than the one- and three-parameter item response models. This finding was not consistent with the earlier findings by Hanson and Gierl (1995). Moreover, this finding is contrary to the common observation that the three-parameter model is the best model for multiple-choice test items (Lord, 1980; Traub, 1983).

Table 14

Comparison of Misfitting Items: TTWI Subtest

	No. of	Conv.		IR		
Model	Items	p-value	pbs	one-	two-	three-
Conv.						
p-value	5	-	1	4	0	2
pbs	2		-	1	1	0
IR						
one-	17			-	4	8
two-	6				-	5
three-	10					-

Comparison of the Ability Estimates

Results of the comparisons of ability estimates for the TTWI subtest are reported in Table 15. The ability estimates were computed using the scoring algorithm associated with the number right and the scoring algorithms associated with each of the item response models and the finite state score model. The correlations among the ability estimates, transformed to a common scale ($\mu=50$, $\sigma=10$), are reported in the upper triangle and the mean absolute differences are reported in the lower triangle.

Table 15

Correlations and Mean Absolute Differences: TTWI Subtest

	NR	θ_1	θ_2	θ_3	λ
NR	-	.999	.968	.911	.923
θ_1	.33 (.38)	-	.969	.908	.932
θ_2	2.01 (1.57)	1.98 (1.50)	-	.938	.899
θ_3	3.16 (2.78)	3.21 (2.85)	2.41 (2.58)	-	.845
λ	2.43 (3.25)	2.13 (3.19)	2.83 (3.66)	3.94 (4.14)	-

Note: NR is the number right score,θ_i indicates ability estimates from item response theory, *i* indicates the number of parameters in the model, and λ is the ability estimate from finite state score theory.

The correlation between the number right scores and the one-parameter ability estimates is nearly perfect. The correlations between the two-parameter ability estimates and each of the number right and one-parameter ability estimates are next in value (.968 and .969). However, the correlations between the three-parameter ability estimates and these estimates are somewhat lower (.911, .908, and .938) as are the correlations between the finite state estimates and the number right, one-, and two-parameter ability estimates (.923, .932, and .899). Further, the correlation between the three-parameter and finite state score estimates is the lowest, .845.

The means of the absolute differences among the pairs of transformed scores, presented in the lower triangle of Table 15, ranged from .33 to 3.94, and followed a pattern consistent with that for the correlations. The closest agreement was between the number right and one-parameter ability estimates (.33). The next small differences, 2.01, and 1.98 were between the two-parameter ability estimates and the number right and one-parameter ability estimates. Both the three-parameter ability estimates and finite state score estimates agreed less closely with ability estimates yielded by number right, one-, and two-parameter models (3.16, 3.21, 2.41; 2.43, 2.10, and 2.83), respectively, and particularly with each other, 3.94.

The standard deviations covaried with the mean absolute deviations in the case of pairs formed from the number right, and the one-, two-, and three-parameter ability estimates. The pattern is broken when the finite state ability estimates became one of the members of a pair. As shown, the variabilities in this case are somewhat larger than those when the finite state ability estimates are not included as a member of a pair. In case of the mean absolute differences, this was not always the case (see Table 15).

In an attempt to explain the larger standard deviations of the absolute differences associated with the pairs involving the finite state ability estimates, the finite state scoring algorithm was examined. This examination revealed that a score of zero was assigned to all examinees who scored at or below the chance level (see Appendix A, lines 3 to 6). In the case of the English subtests reported in the previous chapter, the number of examinees who scored at or below the chance level was 18; 3 for the ENGI subtest and 18 for the ENGDB subtest. In the case of theTTWI subtest, this number was 18. Given that a comparison was to be made between the agreement indices for the TTWI subtest with the subtest containing 18 Test of Test-Wiseness items which violated the option independence assumption, a constant sample size was desirable. The number of examinees who scored at or below chance level on the subtest of 18 items was 57. Accounting for the overlap between the two guessing scores, a total number of 71 examinees was identified; this number represents 5.8% of the total sample. A question then arose as to whether this relatively large number of zero scores led to the low correlations and higher mean absolute differences reported in Table 15. Thus, 71 examinees were removed from the sample. The analyses were then repeated for the reduced sample of subjects (n = 1,161).

In the case of the conventional item analysis, the number of misfitting items using the p-value standard changed from 5 to 2 while using the point-biserial standard the number changed from 2 to 7. However, of these 7 items, 6 had point-biserial values of .19, which is only .01 below the standard value of .20. Further, the changes in the values of these coefficients and the remaining coefficients were within the range of .03. With the item response models, there was a better fit of items to two of the models than when all examinees were included in the analyses; the number of misfitting items decreased from 17 to 12 for the one-parameter model and 10 to 8 for the three-parameter model. The number remained the same, 6 items, for the two-parameter model. The new item parameters based on the 1,161 subjects are presented in Appendix F.

The assessment of the assumption of unidimensionality for the sample of 1,161 subjects revealed that this assumption is still tenuous. The percentage of variance accounted for by the first component 7.4%, was close to the variance accounted for by the second component, 4.4%. The differences between successive components were very small. The scree test indicated the possibility of more than one dimension. However, the Stout's T statistic was marginally not significant ($p \leq .056$), suggesting that the hypothesis that the TTWI subtest was essentially unidimensional was, perhaps, tenable for the reduced sample of examinees.

The correlations and mean absolute differences for the reduced sample of 1,161 examinees are presented in Table 16. As shown in the upper triangle, the correlations among the number right, one-, and two-parameter ability estimates were essentially unchanged following removal of the 71 examinees (cf, Table 15). However, while still

Table 16

Correlations and Mean Absolute Differences for TTWI: Elimination of Examinees With Chance Score

	NR	θ_1	θ_2	θ_3	λ
NR	-	1.000	.967	.940	.965
θ_1	.09 (.09)	-	.967	.941	.968
θ_2	2.07 (1.55)	2.07 (1.54)	-	.987	.935
θ_3	2.73 (2.11)	2.72 (2.10)	1.16 (1.12)	-	.919
λ	1.82 (1.94)	1.73 (1.88)	2.76 (2.29)	3.12 (2.46)	-

Note: NR is the number right score, λ_i indicates ability estimates from item response theory, *i* indicates the number of parameters in the model, and λ is the ability estimate from finite state score theory.

lower than the number right, one-parameter, and two-parameter correlations, the correlations between these ability estimates determined using the three-parameter and the finite state scoring approaches increased. The correlations between the three-parameter ability estimates and the number right, one-, and two- parameter ability estimates changed from .911 to .940, .908 to .941, and .938 to .987, respectively. In the case of the finite state score estimates, the correlations changed from .923 to .965, .932 to .968, and .899 to .935, respectively, and from .845 to .919 with the three-parameter ability estimates.

The mean absolute deviations and their standard deviations followed, with one exception, a pattern consonant with that observed for correlational changes. The mean and standard deviations of the absolute differences between the number right and the one- and two-parameter ability estimates were essentially unchanged. In contrast, while the correlations between number right and the one-parameter model was unchanged, both the mean absolute deviations and the standard deviation of the absolute differences decreased (from .33 to .09 and .38 to .09, respectively). Similarly, the mean absolute deviations and their corresponding standard deviations decreased somewhat for the pairs of estimates in which the three-parameter estimates (3.16 to 2.73, 3.21 to 2.10 and 2.41 to 1.16) and the finite state estimates (from 2.41 to 1.82, 2.13 to 1.73 and 2.83 to 2.76) were included.

In the case of the three-parameter model, removal of the 71 examinees who scored at or below chance level led to a greater agreement with the ability estimates yielded by the number right scores and the one-, and two-parameter models. These findings suggest that the inclusion of the pseudo-guessing parameter does not adequately take into account examinees who score at or below the chance level. Given that the assessment of the assumption of minimal guessing revealed that guessing was not a factor that influenced performance on this subtest, the reasons why the three-parameter model seemed to misbehave in the presence of large number of low scoring examinees were not apparent.

Despite the removal of the 71 examinees who scored at or below chance level, the mean and standard deviation of the absolute differences were still somewhat high for the pairs involving the three-parameter model and the finite state score model. Again, it is interesting to note that, although the number of misfitting items for the one-parameter model was greater than the number for conventional item analysis and the one-, two-, and three-parameter model, this did not seem to adversely influence the correlations and the mean absolute deviations between the one-parameter ability estimates and estimates from the other models. Hence, it is apparent that lack of fit of items is not an explanation for the low correlations and high mean absolute differences in the case of the three-parameter model.

One possible explanation is the possibility of multidimensionality of the item responses since, as indicated before, the methods used to assess the assumption of unidimensionality led to ambiguous results about the tenability of this assumption. Hence, at this point, the reasons for low correlations and high mean absolute deviations are not apparent.

Test-Wiseness Items With a Pair of Opposite or Similar Options

The results presented in this section are based on the analyses of 18 items in the Test of Test-Wiseness which contained either a pair of options which were similar (ID2) or a pair of options which were opposite to one another (ID3). These items were initially developed to assess students' possession of test-wiseness skills and scored for the test-wise strategy used. For example, if two similar options were avoided, then a score of one was awarded for that item. Similarly, if either of the two opposite options was selected, then a score of one is awarded (Rogers & Wilson, 1993). In this study, however, the items were scored only for the correct answer. In the following discussion, this subtest will be referred to as TTWDOS subtest, where DOS signifies dependency among options brought about by the presence of pairs of opposite or similar options.

Subtest Characteristics

The distribution of number right scores for the TTWDOS subtest is presented in Figure 11. The mean score for this set of items was 8.49 (41.2 %) and the corresponding standard deviation was 2.44 (13.5 %). The internal consistency computed for the set of 18 items using the Cronbach's alpha was .32 and the standard error of measurement was 2.02 (11.2 %). The expected internal consistency of a 48 item test obtained by adding 30 parallel items is .56. This value indicates that the consistency of the items in measuring the test-wiseness construct is low. Such a low value is likely due to the guessing, particularly on the 24 novel items (Rogers & Bateson, 1991a, p. 171) and the fact that items were selected from four subject areas. The skewness coefficient was .08 and the kurtosis coefficient was -.09. These values indicate that the shape of the distribution of scores was essentially normal.

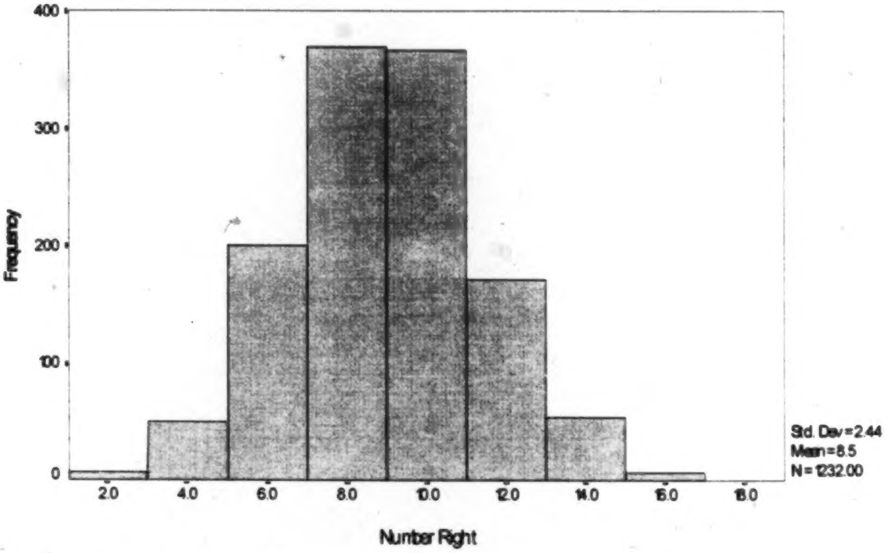


Figure 11: Distribution of Number Right Scores: TTWDOS Subtest

Item Characteristics

The estimated item parameters obtained from the conventional and item response item analyses for each of the 18 items in the TTWDOS subtest are listed in Table 17. In the case of all models, misfitting items are identified by an asterisk (*).

Table 17

Conventional and Item Analysis Item Analysis Results: TTWDOS Subtest

	Conventional			IRT Parameter Estimates					
				1PL	2PL		3PL		
Item	p-value	pbs	option	b	b	a	b	a	c
1	0.67	0.24	✓	-2.15	-2.18	0.19	0.08	0.24	0.34
4	0.33	0.33	b	2.20	1.18	0.39	1.56	0.74	0.00
6	0.40	0.28	✓	1.27	1.00	0.26	1.95	1.05	0.34
10	0.35	0.25	✓	1.85	1.69	0.22	2.84	0.64	0.31
12	0.58	0.24	✓	-0.99	-1.07	0.18	2.84	0.39	0.50
15	0.50	0.31	✓	-0.02	-0.02	0.27	1.70	0.49	0.36
16	0.39	0.28	✓	1.36	1.20	0.22	2.72	0.66	0.34
18	0.67	0.34	✓	-2.16	-1.07	0.43	-0.30	0.58	0.25
20	0.27*	0.25	✓	3.07	2.31	0.27	2.56	0.79	0.22
21	0.58	0.34	✓	-0.97	-0.53	0.39	0.33	0.59	0.26
23	0.29*	0.21	✓	2.70	2.80	0.19	3.61	0.83	0.28
26	0.63	0.23	✓	-1.69	-1.70	0.20	1.08	0.40	0.44
27	0.39	0.35	✓	1.42	0.75	0.39	1.44	0.78	0.25
32	0.42	0.23	✓	0.95	1.03	0.18	2.62	1.07	0.40
34	0.61	0.29	✓	-1.37	-0.92	0.30	0.36	0.46	0.30
37	0.31	0.27	✓	2.50	1.94	0.26	2.11	1.01	0.26
38	0.50	0.35	✓	0.02	0.01	0.37	0.88	0.58	0.26
43	0.59	0.26	✓	-1.16	-1.08	0.21	1.89	0.56	0.50

Conventional Item Analysis Results

The conventional item difficulties, presented in column 2 of Table 17, ranged from .27 to .67. Items 20 (p-value = .27) and 23 (p-value = .29) did not meet the minimum standard of .30. However, none of the items exceeded the maximum standard of .85. Thus, two items failed to meet the first standard for the minimum item difficulty value for the correct option. The values of the point-biserial coefficients ranged from .21 to .35 (see Column 4). Accordingly, all the items met the second standard that the minimum value for item discrimination be at least .20.

An examination of the point-biserials for distractors revealed that of the 54 point-biserials for the distractors, 15 had values between -.05 to -.09 (see Appendix G). The rest had values that were less than or equal to -.10. Only one item, item 4, failed to satisfy the fourth criterion; four percent of the students selected option 3. When all four standards are used, three items failed to meet one of standards set for either the correct option or the distractors.

Item Response Item Analysis Results

Assessing the Assumption of Item Response Theory

Unidimensionality. The eigenvalues and the associated percentages of variance accounted for by the first five of the components extracted from the correlation matrix containing phi coefficients and the correlation matrix containing tetrachoric coefficients are reported in Table 18.

The percentages of variance accounted for by the first components are small for both the phi and tetrachoric correlations (8.5% and 10.4% respectively). However, the changes in the percentages of variance accounted for by the first and second components are somewhat larger than the remaining differences. Thus, following Huynn and Fererra (1994), it appears that the TTWDOS subtest is essentially unidimensional. This finding

Table 18

The First 5 Components: TTWDOS Subtest

	Phi Correlations			Tetrachoric Correlations		
Comp.	λ	% of Var.	% Δ	λ	% of Var.	% Δ
1	1.533	8.5	-	1.872	10.4	-
2	1.219	6.8	1.7	1.358	7.5	2.9
3	1.149	6.4	0.4	1.243	6.9	0.6
4	1.115	6.2	0.2	1.190	6.6	0.3
5	1.094	6.1	0.1	1.158	6.4	0.2

Note: Var. indicates the variance accounted for by the component extracted.
Δ indicates change in the percentage of variances.

is supported by the scree plot; while the scree plot shows the possible existence of a second component, there still appears to be a dominant factor (see Figure 12). Stout's T statistic computed for this set of data was .323 with the *p* value of .373. Therefore, the null hypothesis that $d_E = 1$ was not rejected, implying the tenability of the assumption of essential unidimensionality for the TTWDOS subtest. However, it is pertinent to note that the minimum number of items required for a test to be analyzed using DIMTEST program is 20. Thus, it is possible that failure to meet the minimum number of items required may have affected the value of Stout's T statistic.

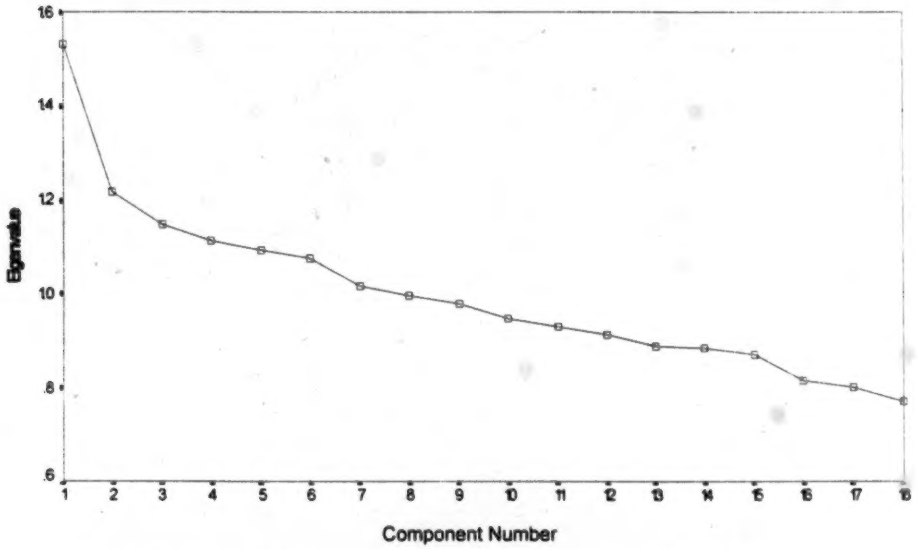


Figure 12: Eigenvalues Against Component Number: TTWDOS Subtest

Equal item discrimination. As shown in column 4 of Table 17, the values of the point-biserial correlations ranged from .21 to .35. It was concluded that the assumption of equal discrimination was met by the TTWDOS subtest because that range falls within the .15 range suggested by Hambleton and Murray (1983).

Nonspeededness. To determine whether or not the examination was speeded, the percentage of examinees who did not complete the test items was examined. Of the 1,232 examinees in the sample, only 2 did not answer one of the last three items in the TTWDOS subtest. Thus, it was concluded that speededness was not a factor that affected the examinees' performances.

Minimal guessing. The performance of low-scoring examinees on the most difficult items, items 20 (p-value = .27), 23 (p-value = .29), and 37 (p-value = .31), was examined with an expectation that these examinees would not be able to answer correctly these items. It was found that 96 (67.1%) of the 143 examinees, who were identified as low-scoring examinees for this subtest had a total score of 0 on these items, 38 (26.6%) had a score of 1, and 9 (6.3%) had a score of 2. Since there were not enough examinees with scores of 1, 2, and 3, it was concluded that the influence of guessing was minimal for the TTWDOS subtest.

Examination of Fit of Items to the Item Response Models

The fit of items to the item response models was assessed using the standardized root mean square residual since the TTWDOS subtest contained 18 items. Mislevy and Bock (1990) suggested that standardized root mean squares greater than 2 be taken as an indication of misfit of an item to the model. Applying this criterion, no items in the TTWDOS subtest were identified as misfitting. The standardized root mean square residual for each of the 18 items was less than 2. This finding indicates that each item fit each of the item response models.

Comparison of Misfitting Items Identified by Conventional and Item Response Item Analyses

Only 2 items were identified as misfitting. Both items failed to meet the minimum difficulty standard set for the conventional item analysis.

Comparison of the Ability Estimates

Results of the comparisons of the ability estimates and the mean absolute differences among the transformed estimates are reported in Table 19 for the 1,232 examinees. As shown in the upper triangle, the correlation between the number right scores and one-parameter ability estimates is nearly perfect (.997). The correlations between the two-parameter ability estimates and the number right and one-parameter ability estimates are next in value .969, and .966. On the other hand, the correlation between the three-parameter ability estimates and these estimates are somewhat lower

Table 19

Correlations and Mean Absolute Differences Among Ability Estimates: TTWDOS Subtest

	NR	θ_1	θ_2	θ_3	λ
NR	-	.997	.969	.847	.847
θ_1	.27 (.71)	-	.966	.849	.847
θ_2	1.98 (1.51)	2.06 (1.58)	-	.902	.840
θ_3	4.29 (3.50)	4.23 (3.51)	3.29 (2.95)	-	.711
λ	3.98 (3.78)	4.02 (3.77)	4.00 (3.95)	5.39 (5.32)	-

Note: NR is the number right score, θ_i indicates ability estimates from item response theory, *i* indicates the number of parameters in the model, and λ is the ability estimate from finite state score theory.

(.847, .849, and .902). Similarly, the correlations between the finite state score estimates and the number right, one-, and two-parameter ability estimates are lower (.847, .847, and .840). The lowest correlation was between the finite state score and the three-parameter ability estimates, .711.

The pattern for mean absolute differences among the pairs of transformed scores is similar to that for correlations. The closest agreement was between the number right and one-parameter ability estimates (.27); the next closest agreements were between the two-parameter estimates and the number right and the one-parameter ability estimates: 1.98 and 2.06. As in the case for correlations, the finite state scores and three-parameter ability estimates agreed less closely with the ability estimates yielded by these models (3.29 to 4.29), and least closely with each other, 5.39.

The standard deviations revealed a pattern similar to that revealed by the mean absolute deviations in the case of pairs formed from the number right, and the one-, two-, and three-parameter ability estimates. Again, this pattern was broken when the finite state ability estimates were one of the members of a pair. As shown, the variabilities in this case are somewhat larger when the finite state ability estimates are included than when they are not included in the pair of estimates compared (see Table 19).

The higher means and larger standard deviations of the absolute differences, particularly for the finite state scoring model, may be attributable to the assignment of zero score to all subjects who scored at or below chance level. Consequently, as with the previous subtest, the examinees who scored at or below the chance level were eliminated and the analyses repeated.

The tests for the assumptions of unidimensionality, equal item discrimination, speededness, and guessing revealed results comparable to those obtained for the full sample of 1,232 examinees. The conventional item analysis revealed that two items failed to meet the standards set: one item was too difficult; the point-biserial for the second item was less than .20. No items were identified as misfitting the item response models (see Appendix H).

Table 20

Correlations and Mean Absolute Differences for TTWDOS: Elimination of Chance Scores

	NR	θ_1	θ_2	θ_3	λ
NR	-	1.000	.960	.839	.951
θ_1	.04 (.03)	-	.960	.839	.951
θ_2	2.26 (1.73)	2.26 (1.73)	-	.927	.912
θ_3	4.48 (3.48)	4.47 (3.47)	3.06 (2.30)	-	.839
λ	2.42 (1.94)	2.41 (1.94)	3.23 (2.62)	4.38 (3.40)	-

Note: NR is the number right score, θ_i indicates ability estimates from item response theory, *i* indicates the number of parameters in the model, and λ is the ability estimate from finite state score theory.

The comparisons among reestimated ability estimates are summarized in Table 20. The correlations between the number right, one-, and two-parameter ability estimates were essentially unchanged after removal of the 71 examinees who scored at or below chance level (cf, Table 19). As in the case of TTWI subtest, while still lower than the number right, one-parameter, and two-parameter correlations, and the correlations between these ability estimates determined using the finite state scoring models increased from .847 to .951; .847 to .951; and .840 to .912. The correlation between the three-parameter ability estimates and finite state scores and two-parameter ability estimates increased from .711 to 839 and .902 to .927 respectively. On the other hand, the correlations between the three-parameter ability estimates and the number right and one-parameter ability estimates decreased from .847 to .839 and .849 to .839.

The mean and standard deviations of the absolute differences between the two-parameter ability estimates and the number right and one-parameter estimates were essentially unchanged. In contrast, while the correlations between number right and the one-parameter model was unchanged, both the mean absolute deviations and the standard deviation of the absolute differences decreased (from .27 to .04 and .71 to .03, respectively). In case of the three-parameter ability estimates, the mean absolute deviations and their corresponding standard deviations remained essentially the same while the means and standard deviations for the pairs of estimates in which the finite state score estimates was included decreased (3.98 to 2.42; 4.02 to 2.41; 4.00 to 3.23; and 5.39 to 4.38 (for means) and 3.78 to 1.94; 3.77 to 1.94; 3.95 to 2.62 and 5.32 to 3.40 (for standard deviations)).

Unlike the TTWI subtest, removal of 71 examinees who scored at or below chance level did not lead to a greater agreement between the three-parameter ability estimates and with the ability estimates yielded by the number right scores and the two-parameter models. Given that the assessment of the assumption of minimal guessing revealed that guessing was not a factor that influenced performance on this subtest, the reasons why the three-parameter ability estimates had a poor agreement with other ability estimates were not apparent.

Despite the removal of 71 examinees who scored at or below chance level, the mean and standard deviation of the absolute differences were still high for the pairs involving the three-parameter model and the finite state score model and the two-parameter model. Only two items failed to meet the standards set for conventional item analysis and no item was identified as misfitting any of the item response models. Hence, lack of fit of items is not an explanation for the low correlations and high mean absolute differences in the case of the three-parameter.

Discussion

Before discussing the comparative findings it is first necessary to address the issue of the assignment of zero score in the finite state scoring algorithm to examinees who scored at or below chance level. As expected, removal of these examinees improved the agreement between the re-estimated finite state ability estimates and the ability estimates yielded by the number right scoring aqlgotithm and one-, two- and three-parameter item response scoring algorithms. However, examinees who score at or below the chance level would not be excluded from the analyses based simply on the test scores. In the case of Alberta Education, the agency which provided the data set, examinees are excluded only when there is evidence of cheating or exceptions in test administration.

Comparison of the correlations between ability estimates yielded by the number right and the one- and two-parameter item response scoring algorithms for the TTWDOS and TTWI subtests reveal that they are essentially the same (cf, Table 15 and 19). In contrast, the correlations in which the three-parameter or the finite state score ability estimates are correlated with these estimates are larger for the TTWI subtest than for the TTWDOS subtest, .911 vs .847; .908 vs .849; and .938 vs .902 (three-parameter estimates); .923 vs .847; .899 vs .840; and .845 vs .711 (finite state score estimates).

The pattern for the differences between the means of the absolute differences for the TTWDOS subtest and the corresponding means for the TTWI subtest is similar to that for correlations (cf, Table 15 and 19). The two sets of mean absolute difference computed from the number right and the one- and two-parameter ability estimates are approximately equal. In contrast, the mean absolute deviations computed from the three-parameter and finite state score ability estimates are smaller for the TTWI subtest than for the TTWDOS subtest, 3.16 vs 4.29; 3.21 vs 4.23; and 2.41 vs 3.29 (three-parameter ability estimates) and 2.43 vs 3.98; 2.13 vs 4.02; 2.83 vs 4.00; and 3.94 vs 5.39 (finite state score ability estimates). With the exception of the number right and one-parameter ability estimates, the pattern for the standard deviations of the absolute differences is the same as that observed for the mean absolute deviations, both within and across the TTWI and TTWDOS subtests. As foreshadowed by Garcia-Perez and Frary (1989), the finite state scoring algorithm was quite sensitive to the violation of the assumption of independent optionsdue to the inclusion of a pair of similar or opposite options.

The presence of items which failed to meet standards set for conventional item analysis and items which were identified as misfitting one or more of the item response models did not seem to adversely influence the correlations, means, and standard deviations of the absolute differences among the apirs of ability estimates. Therefore, it seems that lack of fit of of items is not an explanation for low correlations and high means and standard deviations of the absolute differences.

Another plausible reason why the three-parameter model perfomed poorly is the possible failure of the psuedo-guessing parameter to adequately account for guessing. On the one hand, the test for guessing indicated that the influence of guessing was minimal. in contrast Rogers and Bateson (1991a) found that the students often guessed after eliminating one or two options. Given this latter finding, the psuedo-guessing parameter may not be sensitive enough to the presence of test-wise cues and the use of these cues by the examinees.

Further, it might be asked if the differences observed between the results for the TTWI and the TTWDOS subtest were due to the different number of items, 32 and 18, included in each subtest. Again if this was a factor, all the results should have been affected. But all were not. The poorer performances of the three-parameter and finite state models were not attributable to attenuation.

CHAPTER 6: SUMMARY, CONCLUSIONS AND IMPLICATIONS

This chapter begins with a brief summary of the study, followed by identification of the limitations of the study and the conclusions drawn in light of these limitations. The chapter concludes with implications for practice and for future research in the scoring of multiple-choice test items.

Summary of the Study

Background to the Study

The oldest form of scoring multiple-choice test items involves awarding one point if the correct or keyed option is selected by an examinee and no points if any of the remaining options are selected. However, this number right scoring procedure has been criticized (e.g., Choppin, 1983; Shepard, 1993; Yamagishi, 1991). The principal criticism is centred upon the different ways in which examinees select the correct option and the failure of the number right scoring to capture these differences. More specifically, number right scoring cannot differentiate among examinees who select the correct option because they have knowledge of the correct answer, partial knowledge, or they are simply guessing. Consequently, several alternative scoring methods that take into account such factors have been developed.

The finite state score model (García-Peréz, 1987; García-Peréz & Frary, 1989) is a recently developed alternative scoring method. Congruent with the current interest in merging psychometric methods with the cognitive science (Snow & Lohman, 1989, 1993), finite state score theory attempts to take into account the cognitive processes involved in answering multiple-choice test items, with particular reference to partial knowledge and guessing.

The present study was motivated by the need for research involving the application of the finite state score theory in scoring multiple-choice test items in order to assess its utility in actual testing practices. Except for the studies by Zin (1992) and Ndalichako and Rogers (1997), previous studies in which the finite state score theory has been assessed involved the use of simulated data sets.

Purpose of the Study

The main objective of this study was to compare the ability estimates obtained using the finite state scoring algorithm with the ability estimates obtained from the number right and the item response scoring algorithms in an actual large scale testing program. This comparison was made by examining the similarity of examinees rankings based on their scores and the extent to which the scores agreed with each other.

Tests included in large testing programs contain items which violate one of the principal assumptions underlying the use of finite state scoring, namely the assumption of option independence (García-Peréz and Frary, 1989). For example, an item may call for the best rather than the correct option among the options presented. Other items may contain test-wise elements (Millman, 1966; Rogers & Bateson, 1991; Sarnacki, 1979) such as a pair of opposite options or a pair of similar options. Therefore, the present study explored whether the finite state scoring model would be sensitive to such violations and provide scores which would differ more from scores provided by the number right scoring and by the item response models in tests composed of items which do not satisfy the option independence assumption than in tests composed of items which satisfy this assumption.

In an attempt to explain differences among the models, should they be found, item analyses were completed. Unfortunately, the finite state score model does not provide item level information. Instead, conventional and item response approaches were used. Both approaches were considered because it was not clear whether they would lead to the same decision about the item quality (Knodel, 1981).

Method of Study

Data Analyzed

Two sets of data were analyzed. The first set consisted of responses of 1,232 examinees to the English 30 Diploma Examination administered in the province of Alberta. The second set consisted of responses of the same examinees to the Test of Test-Wiseness (Rogers & Wilson, 1993). Each examination was divided into two subtests. In case of the English 30 Examination, the first subtest, labelled ENGI, consisted of 48 items which satisfied the assumption of option independence; the second subtest (ENGDB) consisted of 20 items which called for the best answer and, therefore, did not satisfy the assumption of option independence. Similarly, the first subtest of the Test of Test-Wiseness, labelled TTWI, consisted of 32 items which satisfied the assumption of option independence and the second subtest (TTWDOS) consisted of 18 items which contained either a pair of opposite or similar options and, therefore, did not satisfy the assumption of option independence.

Analyses Conducted

The item analysis evaluation criteria used by Alberta Education were used in this study to evaluate items in conventional item analysis. Both corrected and uncorrected point-biserial coefficients were used for the two subtests of the English 30 Examination. The uncorrected point-biserial coefficient was used for the two subtests of the Test of Test-Wiseness because this was the discrimination index used by the author of the test (Rogers & Bateson, 1991b). In the case of the item response models, the chi-square statistic was used for the ENGI, ENGDB, and TTWI subtests. The root mean square of the standardized residuals was used for the TTWDOS subtest because the number of items was less than the minimum number of 20 required for the chi-square statistic (Mislevy & Bock, 1990). The infit and outfit statistics were used with the partial credit model, which was used with the ENGDB subtest. The assumptions underlying the item response models -- unidimensionality, equal item discrimination, nonspeededness, and minimal guessing -- were examined prior to conducting the item response item analyses.

Ability estimates obtained from each of the models were then compared in terms of the extent to which (a) the ability estimates provided similar rankings of examinees and (b) the values of the ability estimates agreed with each other. The ability estimates were first transformed to T-scores ($\mu = 50$, $\sigma = 10$) to achieve a common scale. Correlations among the ability estimates were computed for rank comparisons and the mean and standard absolute deviations of the differences between the ability estimates were computed to assess the degree of agreement.

Results and Discussion

Assumptions of Item Response Theory

Unidimensionality. For the English subtests, the percentage of variance accounted for by the first component, comparison of successive differences, the shape of the scree plot, and Stout's T statistic showed the existence of essentially a dominant component underlying the item responses.

In the case of the TTWI subtest, the percentage of variance accounted for by the first component was larger than the percentage of variance accounted for by the second component. Further, differences between successive components beginning with second component were small. These findings indicated the existence of a dominant component. While the shape of the scree plot indicated the possible existence of a second component, it supported the existence of a dominant component. However, the value of Stout's T statistic indicated that the item responses were not essentially unidimensional.

For the TTWDOS subtest, while the percentage of variance accounted for by the first component was somewhat less than for the other subtests, the small changes in the variance accounted for by successive components beginning with component two suggested that the first component was a dominant component. This was also seen in the shape of the scree plot. Further, Stout's T statistic showed that the items were essentially unidimensional. Thus, while the assumption of unidimensionality was not tenable for the TTWI subtest, the assumption was met for the TTWDOS subtest.

Equal item discrimination. The assumption of equal item discrimination was met only by the TTWDOS subtest of 18 Test of Test-Wiseness items. In case of the remaining subtests, there were large variations among the uncorrected point-biserial correlation coefficients for the correct option. Consequently, the assumption of equal item discrimination was not met.

Nonspeededness. There were no evidence of speededness for each of the four subtests.

Minimal guessing. The influence of guessing on the performance of examinees was found to be minimal in case of each subtest.

Examination of the Fit of the Items

English subtest. With the exception of the one-parameter model, the number of items identified as misfitting was less than or equal to 12.5% for the ENGI subtest and 20% for the ENGDB subtest. For the one-parameter model, the percentages were greater (43.7% and 60%). This was likely because of the lack of homogeneity of item discrimination.

Test of Test-Wiseness. In contrast to the ENGDB subtest, only 2 items failed to meet the evaluation criteria set for the conventional item analysis in the case of TTWDOS. With the TTWI subtest, the number of misfitting items was greater in the case of the one-parameter model (53.1%). For other models, the number of misfitting items was at most 31.2%. On the surface, the relative greater number of items identified as misfitting the item response models in the case of TTWI can be attributed to the multidimensional structure of the Test of Test-Wiseness. Items included represent four subject areas. However, this seems not to be the case because only two items were identified as misfitting in the case of TTWDOS.

It is evident that different conclusions about the fit of an item may be drawn using different item analyses approaches. Items which misfit an item response model may not fail the standards set in a conventional item analysis. Similarly, depending on which item response model is used, the decision about the item quality could be different. In the light of such inconsistency, it was difficult to identify which model provided the right or correct results.

Concerns about the effectiveness of fit statistics were raised by Traub (1983). He noted that "there exists no basis in inductive logic for concluding that a model fits data, yet it is just this conclusion we wish to justify ... tests of fit can only provide means for rejecting the items but not accepting them" (p. 57). Indeed, goodness-of-fit investigations are essentially the assessment of model-data misfit. In contrast, if there were means of determining whether the item is acceptable by focussing on the values of its parameters, as in the case of the conventional item analysis, a better understanding of how the items function might be reached than when other statistics are used.

Comparison of the Ability Estimates

English subtests. The correlations between the number right, one-, two-, and three-parameter ability estimates were quite high (≥ .98) for both the ENGI and ENGDB subtests. These findings indicate that there are essentially no differences in the extent to which the ability estimates derived from the number right scoring algorithm and scoring algorithms associated with the item response models ranked examinees. However, the correlations between the finite state ability estimates and the ability estimates yielded by the one-, two-, and three-parameter scoring algorithms dropped slightly with the introduction of the best answer items. Moving from the ENGI to the ENGDB subtest, the correlations changed from .984 to .963, .981 to .957, and .970 to .952, respectively. The correlations between the finite state and number right ability estimates remained constant (.957 vs .955).

The mean absolute deviations followed a pattern similar to the pattern for the correlation coefficients. The mean absolute deviations in which the finite state scores were one of the two estimates being compared were (a) for both the ENGI and ENGDB subtests, greater than the mean absolute deviations in which pairs involving the number right, and the one-, two-, and three-parameter ability estimates were compared and, with one exception, (b) greater in the ENGDB subtest than the ENGI subtest (2.15 vs 2.22; 1.81 vs 1.36; 2.01 vs 1.36; and 2.10 vs 1.78; respectively). Likewise, the variability of the absolute deviations which included the finite state ability estimates were (a) for both subtests, greater than the absolute deviations for the pairs in which the number right, and the one-, two-, and three-parameter ability estimates were compared and, with one exception, (b) greater in the ENGDB subtest than the ENGI subtest (1.99 vs 2.02; 1.96 vs 1.46; and 2.13 vs 1.77, respectively).

The partial credit model behaved somewhat differently; the correlations for the pairs involving partial credit ability estimates were systematically lower than the correlations among the number right, one-, two-, and three-parameter, and finite state ability estimates. Likewise, the corresponding means and standard deviations of the absolute differences involving the partial credit ability estimates were somewhat greater than the mean and standard deviations of the absolute differences between the pairs of number right, one-, two-, and three-parameter, and finite state ability estimates.

Despite being in the expected direction, the differences between the finite state ability estimates and the number right, one-, two-, and three-parameter models were not as large as expected. As pointed out in Chapter 4, an order from best to least best option for the best answer items was problematic. There was evidence to suggest that, where ordering did not exist, it existed only for the best and second best options. Consequently, it may be that as a set, the 20 best answer items were more like "modified correct answer" items than “full” best answer items. Hence, the options in many of these items were options for which an independent truth value of the options could be assigned. Therefore, the assumption of option independence would have been partially met. Given this condition, the apparent failure of the partial credit model is attributable to misapplication of this model; the items were not fully partial credit items. In the case of finite state scoring model, while the differences were not large, the observation that they were in the expected direction suggest that the finite state scoring model is sensitive to lack of option independence in best answer items.

Test of Test-Wiseness. As pointed out in Chapter 5, the correlations, means and standard deviations of the absolute differences for the pairs of ability estimates in which the finite state ability estimates were included were, respectively, somewhat lower and higher than expected, given the correlations and means and standard deviations for the other pairs of ability estimates. Examination of the scoring algorithm for the finite state model revealed that examinees who score at or below the chance level are assigned a value of zero. Seventy one examinees, representing 5.8% of the total sample, scored at or below the chance level and were, therefore, assigned a score of zero. Removal of these examinees led to some improvement in the agreement among the ability estimates when the finite state scores were one of the estimates being compared. However, test agencies do not typically exclude examinees for reasons other than cheating and exceptions in test administration. Consequently, the results reported in this section were based on the full sample of 1,232 examinees.

For both the TTWI and TTWDOS subtest, the correlations between the number right, one-, and two-parameter ability estimates were quite high (≥ .97). Again, these findings indicate that there are essentially no differences in the extent to which the ability estimates yielded by the number right scoring algorithm and scoring algorithms associated with the one- and two-parameter item response models ranked examinees. In contrast, the correlations in which the three-parameter or the finite state score ability estimates are correlated with these estimates were lower for the TTWDOS subtest than for the TTWI subtest, .847 vs .911; .908 vs .849; and .938 vs .902 (three-parameter ability estimates); .847 vs .923; .840 vs .899; and .711 vs .845 (finite state score estimates).

The mean absolute deviations revealed a pattern consistent with that revealed by the correlation coefficients. Again, the mean absolute deviations in which the finite state scores or the three-parameter estimates were one of the two estimates being compared were (a) for both subtests, greater than the mean absolute deviations in which the pairs of the number right, and the one-, and two-parameter ability estimates were compared and (b) greater in the case of the TTWDOS subtest than the TTWI subtest. Likewise, the variability of the absolute deviations which included the finite state scores or the three-parameter ability estimates were (a) for both subtests, greater than the standard deviations of the absolute differences for the pairs in which the number right, and one-, and two-parameter ability estimates were compared and (b) greater in the case of the TTWDOS subtest than the TTWI subtest. Again, lack of independence among the options due to the presence of a pair of similar options or a pair of opposite options seemed to adversely influence the ability estimates, for pairs involving the finite state score ability estimates, for the TTWDOS subtest.

Limitations of the study

The principal limitation of the present study was the lack of item level results yielded by the application of the finite state score model. The analyses based on the finite state score theory are only at the test level; analyses at the item level have not yet been incorporated into the theory. This lack of item level information hindered a thorough comparison of the finite state scoring model with the other scoring models. An understanding of item characteristics from the perspective of the finite state scoring model would have helped to explain the discrepancies observed in the correlations, the means, and standard deviations of the absolute differences for the pairs involving the finite state score estimates when the best answer items or items which contained either a pair of opposite options or a pair of similar options were used.

Conclusions and Implications for Practice

The correlations between the finite state ability estimates and number right and one-, two-, and three-parameter ability estimates tended to be comparable to the correlations among these number right and one-, two-, and three-parameter estimates when the assumption of option independence was met. However, they tended to be somewhat lower when the assumption of option independence was violated through the presence of best answer items or items with a pair of opposite options or a pair of similar options. Congruent with this finding were higher means and larger standard deviations of the absolute difference scores when the ability estimates yielded by the finite state scoring model were included in the difference. The higher means imply larger differences between finite state ability estimates and the estimates obtained from the other scoring algorithms; the larger standard deviation indicates that for some students the differences are quite large. Consequently, the use of the finite state scoring procedure instead of one of the other scoring procedures could lead to a different decision.

García-Peréz and Frary (1989) warned that the presence of items which violate the assumption of independence would influence the finite state ability estimates. Given that best answer items are frequently used, and the difficulty of avoiding pairs of similar or opposite options, the use of finite state scoring is problematic in actual testing situations.

Further, the assignment of zero scores to examinees who scored at or below the chance level in the finite state scoring algorithm is questionable and may confound the proper interpretation of the scores obtained. The *Principles for Fair Student Assessment Practices for Education in Canada (1993)* caution users of test scores to "avoid misinterpreting scores on the basis of unjustified assumptions about the scoring system used" (p. 18). The interpretation of each examinee's chance score as a zero score is a controversial issue in the measurement field (Poizner, Nicewander, & Gettys, 1978). The *Principles* further suggest that in assessments involving selection items, the directions should encourage students to answer all items without threat of penalty; the use of correction formula is discouraged (p. 8). Further, when students are expected to answer test items, they have the right to know, before they respond, how their responses are going to be scored since the scoring system may have some impact on the way they respond. This point is well advocated in the *Principles for Fair Student Assessment Practices for Education in Canada (1993):*

> Before an assessment method is used, students should be told how their responses or the information they provide will be judged or scored. Informing students prior to the use of assessment method about scoring procedures to be followed should ensure that similar expectations are held by both students and their teachers. (p. 9)

It is not easy to explain to examinees how their scores are derived from the finite state score model. The ability estimates computed from this model are not just the sum of the points accumulated by answering the item correctly. Rather, the total score or ability estimates are obtained through the use of mathematical models which are not simply the linear transformation of the sum of the individual item scores. Given the complexity of this scoring algorithm and the estimation procedure, it is not possible for the classroom teachers and their students to verify the accuracy of the estimates and make independent judgements about the reasonableness of the results obtained. Consequently, given the inferior results when the finite state scoring model is compared with other models, use of the current scoring algorithm and hence the finite state score model is, at this time, not supportable.

A question that arises with the use of conventional item analysis associated with number right scoring is its sample dependency. While this might influence the item characteristics when a sample of examinees tested does not adequately represent the population of interest, in testing programs where all students are assessed, such an influence is debatable. Therefore, in light of the simplicity of the number right scoring and conventional item analysis based upon p-values and point-biserials or discrimination (D) indices, it is concluded that for purposes of scoring and analyzing test items, the conventional approach should be used.

However, three needs that often arise in large scale testing programs are the need to equate multiple forms of a test, measure change from one occasion to another (e.g., from one year to the next) and detect differential item functioning. Although procedures based on elementary statistics (Angoff, 1971; Holland & Thayer, 1988) may be used for these analyses, item response models may provide some additional advantages (Cole & Moss, 1989). However, despite the apparent advantages of the item response models in these situations, many testing agencies like the American College Testing Program (e.g., Cope, 1989) and Educational Testing Service (e.g., Lawrence & Dorans, 1988) use several of these procedures then examine the results for convergence. Therefore, when the analyses are intended for global issues, both the conventional and item response approaches may be used. Further, use of a combination of approaches might lead to a more thorough understanding of such issues.

Turning to the classroom level, it is important to note that the application of the item response models requires large sample size for estimation purposes. While such samples are available in large scale testing programs, they are unlikely to be available at the classroom level. Further, classroom tests are likely to contain, proportionally, a large number of items with test-wise cues like pairs of similar or opposite options (e.g., Hughes, Salvia, & Bott; 1991; Rogers & Bateson, 1991b). Consequently, the use of the finite state score model is not recommended; number right scoring and conventional item analysis are the most appropriate approaches for classroom testing.

Implications for Future Research

Given the findings presented, it is suggested that scoring algorithm for the finite sate score model be modified in such a way that it will provide item level information. Item analysis is an essential ingredient that enables test developers and test users to determine the functioning of the items. Apparently, this feature is lost when finite state score model is used as a method of scoring test items.

Furthermore, the approach of dealing with examinees who score at or below chance level needs to be revised to allow differentiation even for low scoring examinees. Given the findings of this study, when a large number of examinees score at or below chance level, the interpretation of their ability based on such scores may be confounded. Consequently, a proper way of dealing with such scores is desirable in the case of the finite state scoring algorithm.

The three-parameter model misbehaved in the case where there was a substantially large number of low scoring examinees. Given that the three-parameter model incorporates a guessing parameter, such misbehaviour was unexpected and the possible reasons for such a finding could not be established. Therefore, the performance of the three-parameter model needs to be further examined in the case where a large number of examinees have low scores.

The findings of this study revealed poor performance of the partial credit model. While this finding was attributable to the lack of options that were all rankable in terms of the degree of correctness, a protocol analysis needs to be conducted to confirm whether examinees are indeed attempting to rank the options when responding to the best answer items.

Only one subject area, English 30 Reading Comprehension, and a Test of Test-Wiseness composed of items from four subject areas were analyzed in this study. Further studies conducted across different subject areas would help to clarify the generalizability of the findings of this study.

References

Ackerman, T. (1989). Unidimensional item response theory calibrations of compensatory and nonconcompensatory multidimensional items. Applied Psychological Measurement, 13, 113-127.

Aghbar, A. J., & Tang, H. (1991). Partial credit scoring of cloze-type items. (ERIC Document Reproduction Service No. ED 339 201)

Aiken, L. R. (1987). Testing with multiple choice items. Journal of Research and Development in Education, 20(4), 44-58.

Alberta Education (1992). 1992-93 English 30 information bulletin: Diploma Examinations program. Edmonton, AB: Author.

Allen, M. J., & Yen, W. M. (1979). Introduction to measurement theory. Monterey: Brooks/ Cole Publishing Company.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (ed.), Educational measurement (2nd edition) (pp 508-600).

Angoff, W. H. (1989). Does guessing really help? Journal of Educational Measurement, 26, 323-336.

Angoff, W. H., & Schrader, B. W. (1986). A study of hypotheses basic to use of rights and formula scores. Journal of Educational Measurement, 21, 1-17.

Ary, D., Jacobs, L. C., & Razavieh, A. (1996). Introduction to research in education (5th edition). New York: Holt Rinehart and Winston.

Assessment System Corporation (1993). User's manual for ITEMAN: Conventional item analysis program. St. Paul, MN: Author.

Batley, M., & Boss, M. W. (1993). The effects on parameter estimation of correlated dimensions and a distribution-restricted trait in multidimensional item response model. Applied Psychological Measurement, 17, 131-141.

Bennet, R. E., & Ward, W. C. (1993). Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment. New Jersey: Lawrence Erlbaum Associates.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), Statistical theories of mental test scores (pp. 395-479). Reading, MA: Addison-Wesley.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. Psychometrika, 37, 29-51.

Budescu, D., & Bar-Hillel, M. (1993). To guess or not to guess: A decision-theoretic view of formula scoring. Journal of Educational Measurement, 30(4), 277-291.

Cattell, R. B. (1952). Factor analysis. New York, NY: Harper and Bros.

Choppin, B. H. (1983). Extracting more information from multiple-choice tests: Analytic technique for answer-until-correct mode. Paper presented at the Annual meeting of the American Educational Research Association (67th, Montreal, Quebec, April 11-15). (ERIC Document Reproduction Services No. ED 227 175).

Claudy, J. G. (1978). Biserial weights: A new approach to test item option weighting. Applied Psychological Measurement, 2, 25-30.

Cole, N. C., & Moss, P. A. (1989). Bias in test use. In R. L. Linn (Ed.), Educational Measurement (pp 201-220). Washington, DC: American Council on Education and MacMillan Publishing Company.

Cope, R. T. (1989). Application of equipercentile techniques to test scale construction: Scaling and equating of the ACT ASSET Placement Battery. ACT Research Report Series 89-3. Iowa City, Iowa: American College Testing Program.

Crocker, L., & Algina, J. (1986). Introduction to classical and modern test theory. London: Harcourt Brace Jovanovich College Publishers.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), Educational measurement (2nd ed.), (pp. 443-507). Washington, D.C.: American Council on Education.

Davis, F. B., & Fifer, G. (1959). The effects on test reliability and validity of scoring aptitute and achievement tests with weights for every choice. Educational and Psychological Measurement, 19, 159-170.

De Finetti, B. (1965). Methods for discriminating levels of partial knowledge. British Journal of Mathematical and Statistical Psychology, 18, 87-123.

Diamond, J. J., & Evans, W. (1973). The correction for guessing. Review of Educational Research, 43, 181-191.

Dinero, T. E., & Haertel, E. (1977). Applicability of the Rasch model with varying item discriminations. Applied Psychological Measurement, 1, 581-592.

Downey, R. G. (1979). Item-option weighting of achievement tests: Comparative study of methods. Applied Psychological Measurement, 3, 553-461.

Ebel, R. L., & Frisbie, D. A. (1991). Essentials of educational measurement (5th ed.). Englewood Cliffs, NJ: Prentice-Hall.

Feldt, L. S. (1993). The relationship between the distribution of item difficulties and test reliabilty. Applied Measurement in Education, 6(1), 37-48.

Feldt, L. S., & Brennan, R. L. (1989). In R. L, Linn. (Ed.), Educational measurement (3rd ed.), (pp. 105-146). New York: Macmillan.

Frary, R. B. (1988). Formula scoring of multiple choice tests (correction for guessing). Educational measurement: Issues and practices, 7(2), 33-37.

Frary, R. B. (1989). Partial-credit scoring methods for multiple choice tests. Applied Measurement in Education, 2(1), 79-96.

García-Peréz, M. A., (1985). A finite state theory of performance in multiple choice tests. Proceedings of the 16th European Mathematical Psychology Group Meeting, (pp. 55-67). Montpellier, France.

García-Peréz, M. A.(1987). A finite state theory of performance in multiple choice tests. In E. Roskam & R. Suck (Eds.), Progress in mathematical psychology-I, (pp. 455-464). Amsterdam: Elsevier.

García-Peréz, M. A. (1989). Item sampling, guessing, partial information, and decision-making in objective tests: Finite states versus continuous distributions. British Journal of Mathematical and Statistical Psychology, 43, 73-91.

García-Peréz, M. A., & Frary, R. B. (1989). Psychometric properties of finite state scores versus number-correct and formula scores: A simulation study. Applied Psychological Measurement, 13, 403-417.

García-Peréz, M. A., (1990). A comparison of two models of performance in objective tests: Finite state versus continuous distributions. British Journal of Mathematical and Statistical Psychology, 43, 73-91.

García-Peréz, M. A., & Frary, R. B. (1991a). Finite state polynomic item characteristic curves. British Journal of Mathematical and Statistical Psychology, 44, 45-75.

García-Peréz, M. A., & Frary, R. B. (1991b). Testing finite state models of performance in multiple choice tests using items with 'none of the above' as an option. In J. P. Doignon & J. C. Falmagne (Eds.), Mathematical psychology: Current developments, (pp. 273-291). New York: Springer-Verlag.

García-Peréz, M. A., (1993). In defence of none of the above. British Journal of Mathematical and Statistical Psychology, 46, 213-229.

García-Peréz, M. A., (1994). Parameter estimation and goodness-of-fit testing in multinomial models. British Journal of Mathematical and Statistical Psychology-, 47, 247-282.

Gierl, M., & Hanson, A. (1995). Evaluating the goodness-of-fit between Alberta Education Achievement Test data and models assumptions in unidimensional item response theory. (Research Report RR-95-01). Urbana, IL: University of Illinois, Department of Educational Psychology.

Gilman, D. A., & Ferry, P. (1972). Increasing test reliability through self-scoring procedures. Journal of Educational Measurement, 9, 205-207.

Glass, G. V & Hopkins, K. D. (1984). Statistical methods in education and psychology. London: Allyn and Bacon.

Glaser, R. (1981). The future of testing: A research agenda for cognitive psychology and psychometrics. American Psychologist, 36, 923-936.

Green, K. E. (1991). Measurement theory. In K. E. Green (Ed.), Educational Testing: Issues and Applications, (pp. 3-26). New York: Garland Publishing, Inc.

Guliksen, H. (1950). Theory of mental tests. New York: Wiley.

Haladyana, T. M., & Downing, S. M. (1989). A taxonomy of multiple choice item-writing rules. Applied Measurement in Education, 2, 37-50.

Haladyana, T. M. (1990). Effects of empirical option weighting on estimating domain scores and making pass/fail decisions. Applied Measurement in Education, 3(3), 231-244.

Haladyana, T. M. (1994). Developing and validating multiple-choice test items. Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Haladyana, T. M. (1997). Writing test items to evaluate higher order thinking. Boston: Allyn and Bacon.

Hambleton, R. K., Roberts, D. M., & Traub, R. E. (1970). A comparison of reliability and validity of two methods of assessing partial knowledge on multiple choice tests. Journal of Educational Measurement, 7, 75-82.

Hambleton, R. K., & Cook, L. L. (1977). Latent trait models and their use in the analysis of educational test data. Journal of Educational Measurement, 14, 75-96.

Hambleton, R. K., & Murray, L. (1983). Some goodness of fit investigations for item response models. In R. K. Hambleton (Ed.), Applications of item response theory (pp. 71-94). Vancouver, British Columbia, Canada: Educational Research Institute of British Columbia.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of item response theory. Newburg Park, CA: Sage.

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. Educational Measurement: Issues and Practices, 12, 38-46.

Hanna, G. S. (1977). A study of reliability and validity of multiple-choice tests with an answer-until-correct procedure. Journal of Educational Measurement, 14, 1-8.

Huynh, H., & Ferrara, S. (1994). A comparison of equal percentile and partial credit equating for performance-based assessments composed of free-response items. Journal of Educational Measurement, 31, 125-142.

Hughes, C. A., Salvia, J. & Bott, D. (1991). The nature and the extent of test-wiseness cues in seventh- and tenth-grade classroom tests. Diagnostique, 16 (2-3), 153-163.

Holland, P. W., & Thayer, D. T. (1988). Differential item functioning and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), Test validity (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum.

Haney, W. (1982). Trouble over testing. Educational Leadership, 37, 640-650: (ERIC Document Reproduction Service No. ED 205 129).

Hopkins ,K. D., Stanley, J. C., & Hopkins, B. R. (1990). Educational and psychological measurement and evaluation (7th edition.). Englewood Cliffs, NJ: Prentice Hall.

Hutchinson, T. P. (1982). Some theories of performance in multiple choice tests, and their implications for invariant of task. British Journal of Mathematical and Statistical Psychology, 35, 71-89.

Hutchinson, T. P. (1985). Predicting performance in variants of the multiple-choice test. Paper presented at the 4th Annual Meeting of the Psychometric Society and the Classification Societies. Cambridge. (ERIC Document Reproduction Service No. ED 263 177).

Kansup, W. (1973). A Comparison of several methods of assessing partial knowledge in multiple-choice tests. Unpublished masters thesis. University of Alberta.

Knodel, J. W. (1981). A comparison of conventional and Rasch item analysis approaches applied to a grade four science test-pool. Unpublished Masters thesis. University of British Columbia, Department of Educational Psychology.

Lawrence, I. M., & Dorans, N. J. (1988). A comparison of observed score and true score equating methods for representative samples and samples matched on an anchor test. Research Report. New Jersey: Educational Testing Services.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Lord , F. M. (1952). A theory of test scores (Psychometric Monograph, No. 7). Psychometric Society.

Lord , F. M., & Novick, M. R. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

Maguire, T. O., Hattie, J., & Haig, B. (1994). Construct validity and achievement assessment. The Alberta Journal of Educational Research, XL(2), 109-126.

Masters, G. N. (1982). A Rasch model for partial credit scoring. Psychometrika 47(2), 149-173.

Masters, G. N. (1988). The analysis of partial credit scoring. Applied Measurement in Education, 1(4), 279-298.

McKinley, R. L., & Mills, C. C. (1985). A comparison of goodness-of-fit statistics. Applied Psychological Measurement, 9, 49-57.

Mehrens, W. A., & Lehmann, I. J. (1991). Measurement and evaluation in education and psychology. Toronto: Holt, Rinehart, and Winston, Inc.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational Measurement (pp 13-103). Washington, DC: American Council on Education and MacMillan Publishing Company.

Millman, J. (1966). Test-wiseness in taking objective achievement and aptitude examinations. Final Report. New York: College Entrance Examination Board.

Mislevy, R. J., & Bock, R. D. (1990). BILOG 3: Item analysis and test scoring with binary logistic models. Chicago: Scientific Software.

Nandakumar, R. (1993). Assessing essential unidimensionality of real data. Applied Psychological Measurement, 17, 29-38.

Nandakumar, R. (1994). Assessing dimensionality in set of item responses: Comparison of different approaches. Journal of Educational Measurement, 31, 17-35.

Nandakumar, R., & Stout, W. F. (1993). Refinements of Stout's procedure for assessing latent trait dimensionality. Journal of Educational Statistics, 18, 41-68.

Ndalichako, J. L. & Rogers W. T . (1997). Comparison of finite state theory, classical test theory, and item response theory in scoring multiple-choice items. Educational and Psychological Measurement, 57(4), 580-589.

Poizner, S. B., Nicewander, W. A., & Gettys, C. F. (1978). Alternative response scoring methods for multiple-choice items: An empirical study of probabilistic and ordinal response models. Applied Psychological Measurement, 2(1) 83-96.

Pugh, R. C., & Brunza, J. J. (1975). Effects of a confidence weighted scoring system on measures of test reliability and validity. Educational and Psychological Measurement, 35, 73-78.

Rasch, G. (1960). Probabilistic models for some intelligence and attainments tests. Copenhagen: Danish Institute for Educational Research.

Reckase, M. D., & McKinley, R. L. (1991). The discriminating power of items that measure more than one dimension. Applied Psychological Measurement, 15, 361-373.

Rogers, W. T., & Bateson, D. J. (1991a). Verification of a model of test-taking behaviour of high school seniors. Journal of Experimental Education, 59, 331-350.

Rogers, W. T., & Bateson, D. J. (1991b). The influence of test-wiseness on performance of high school seniors on school leaving examinations. Journal of Experimental Education, 59(4), 331-350.

Rogers, W. T., & Wilson, C. (1993). The influence of test-wiseness upon performance of high school students on Alberta Diploma Examinations. Unpublished research report.

Rothman, R. (1995). Measuring up: Standards, assessment, and school reform. San Francisco: Jossey-Bass Publishers.

Rowley, G. L., & Traub, R. E. (1977). Formula scoring, number-right scoring, and test-taking strategy. Journal of Educational Measurement, 14, 15-22.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. Psychometric Monograph, 17.

Sarnacki, R. E. (1979). An examination of Test-Wiseness in the cognitive domain. Review of Educational Research, 49(2), 252-279.

Shepard, L. A. (1993). The place of testing reform in educational reform - A reply to Cizek. Educational Researcher, 22, 10-13.

Smith, R. M. (1987). Assessing Partial knowledge in vocabulary. Journal of Educational Measurement, 24(3), 217-233.

Snow, R. E, & Lohman, D. F. (1989). Implications of cognitive psychology for educational measurement. In R. L. Linn (Ed.), Educational Measurement (3rd ed.) (pp. 263-332). Washington, DC: American Council on Education and MacMillan Publishing Company.

Snow, R. E., & Lohman, D. F. (1993). Cognitive psychology, new test design and new test theory: An introduction. In N. Frederiksen, R. J. Mislevy, & I. I. Bejar (Eds.), Test theory for a new generation of tests (pp. 1-17). Hillsdale, NJ: Lawrence Erlbaum Associates.

Spearman, C. (1904). The proof and measurement of association between two things. American Journal of Psychology, 15, 251-258.

Stanley, J. C. (1971). Reliability. In R. L. Thorndike (Ed.), Educational measurement (2nd ed.) (pp. 356-442). Washington, D.C.: American Council on Education.

Stout, W. F. (1987). A non parametric approach for assessing latent trait unidimensionality. Psychometrika, 52, 589-617.

Stout, W. F. (1990). A new item response theory modelling approach with applications to unidimensionality assessment and ability estimation. Psychometrika, 55, 293-325.

Suen, H. K. (1990). Principles of test theories. Hillsdale, New Jersey: Lawrence Erlbaum Associates, Publishers.

Thissen, D., Steinberg, L., & Fitzpatrick, A. R. (1989). Multiple choice models: The distractors are also part of the item. Journal of Educational Measurement, 26(2) 161-177.

Thorndike, R. M., Cunningham, G. K., Thorndike, R. L., & Hagen, E. P. (1991). Measurement and evaluation in psychology and education. New York: NY Macmillan Publishing Company.

Traub, R. E. (1983). A priori consideration in choosing an item response model. In R. K. Hambleton (Ed.), Applications of item response theory (pp. 57-70). Vancouver, BC: Educational Research Institute of British Columbia.

Traub, R. E. (1994). Reliability for the social sciences: Theory and applications (volume 3). London: Sage.

Wainer, H. (1989). The future of item analysis. Journal of Educational Measurement, 26(2), 191-208.

Wright , B. D., & Masters, G. N.(1982). Rating scale analysis. Chicago: MESA Press.

Wright, B. D., & Linacre, J. M. (1992). A user's guide to BIGSTEPS: Rasch-Model Computer Program, Version 2.2. Chicago: MESA Press.

Yamagishi, M. (1991). Testing in Japan. In K. E. Green (Ed.), Educational Testing: Issues and Applications, (pp. 169-196). New York: Garland Publishing, Inc.

Zin, T. T., & Williams J. (1991). Searching for better scoring of multiple choice tests: Proper treatment of misinformation, guessing and partial knowledge. (ERIC Document Reproduction No. ED 339 744)

Zin, T. T. (1992). Comparing 12 finite state models of examinee performance on multiple-choice tests. Unpublished doctoral dissertation, Virginia Polytechnic and State University.

Appendix A

A Program Used to Compute the Finite State Score Ability Estimates

Appendix A: A Program Used to Compute the Finite State Score Ability Estimates

```
get file= 'data file'
compute ni=nr+nw+no .
compute #check=nw/3.
numeric esti (f8.3) .
do if (nr le #check) .
compute esti=0 .
else if (nr ge ni) .
compute esti=1 .
else .
compute #smin=99999999 .
loop #i=1 to 999 .
compute #w=#i/1000 .
compute #c1=#w**4+4*#w**3*(1-#w)+(9/2)*#w**2*(1-#w)**2 +
2*#w*(1-#w)**3 .
compute #c2=(3/2)*#w**2*(1-#w)**2+2*#w*(1-#w)**3 .
compute #c3=(1-#w)**4 .
compute #chi1=((nr-ni*#c1)**2)/(ni*#c1) .
compute #chi2=((nw-ni*#c2)**2)/(ni*#c2) .
compute #chi3=((no-ni*#c3)**2)/(ni*#c3) .
compute #stat=#chi1+#chi2+#chi3 .
do if (#stat lt #smin).
compute #smin=#stat .
compute esti=#w .
end if .
end loop .
end if .
execute .
list .
```

Appendix B

Conventional Item Analysis: ENGI Subtest

Appendix B: Conventional Item Analysis: ENGI Subtest

		Item Statistics		Option Statistics			
Seq. No.	Scale -Item	Prop. Correct	Point Biser.	Alt.	Prop Endorsing	Point Biser.	Key
1	0-1	.84	.21	1	.04	-.09	
				2	.09	-.13	
				3	.03	-.13	
				4	.84	.21	*
				Other	.00	-.01	
2	0-2	.79	.34	1	.03	-.11	
				2	.79	.34	*
				3	.08	-.20	
				4	.10	-.21	
				Other	.00	-.08	
3	0-3	.66	.45	1	.12	-.20	
				2	.12	-.28	
				3	.66	.45	*
				4	.10	-.19	
				Other	.00	-.04	
4	0-4	.72	.39	1	.10	-.19	
				2	.72	.39	*
				3	.12	-.24	
				4	.06	-.17	
				Other	.00		
6	0-5	.60	.28	1	.60	.28	*
				2	.03	-.14	
				3	.23	-.14	
				4	.14	-.15	
				Other	.00		
7	0-6	.75	.27	1	.09	-.14	
				2	.03	-.16	
				3	.13	-.15	
				4	.75	.27	*
				Other	.00	.01	
9	0-7	.51	.31	1	.25	-.15	
				2	.22	-.17	
				3	.02	-.14	
				4	.51	.31	*
				Other	.00		
10	0-8	.52	.34	1	.29	-.22	
				2	.04	-.10	
				3	.15	-.13	
				4	.52	.34	*
				Other	.00		
11	0-9	.78	.46	1	.78	.46	*
				2	.03	-.14	
				3	.17	-.37	
				4	.02	-.18	
				Other	.00	-.09	

Appendix B Continued

		Item Statistics		Option Statistics			
Seq. No.	Scale -Item	Prop. Correct	Point Biser.	Alt.	Prop Endorsing	Point Biser.	Key
13	0-10	.75	.43	1	.75	.43	*
				2	.11	-.26	
				3	.09	-.22	
				4	.05	-.20	
				Other	.00	-.02	
15	0-11	.83	.30	1	.02	-.10	
				2	.05	-.02	
				3	.83	.30	*
				4	.11	-.31	
				Other	.00		
17	0-12	.77	.34	1	.07	-.16	
				2	.03	-.10	
				3	.77	.34	*
				4	.12	-.26	
				Other	.00	-.02	
19	0-13	.82	.31	1	.08	-.16	
				2	.02	-.08	
				3	.82	.31	*
				4	.08	-.24	
				Other	.00		
20	0-14	.68	.35	1	.10	-.19	
				2	.14	-.18	
				3	.08	-.17	
				4	.68	.35	*
				Other	.00		
23	0-15	.79	.37	1	.03	-.13	
				2	.79	.37	*
				3	.08	-.28	
				4	.09	-.18	
				Other	.00		
24	0-16	.75	.46	1	.07	-.27	
				2	.75	.46	*
				3	.11	-.25	
				4	.08	-.20	
				Other	.00		
25	0-17	.40	.42	1	.05	-.13	
				2	.27	-.14	
				3	.28	-.25	
				4	.40	.42	*
				Other	.00		
26	0-18	.50	.33	1	.12	-.10	
				2	.50	.33	*
				3	.13	-.22	
				4	.25	-.14	
				Other	.00	-.03	

Appendix B Continued

		Item Statistics		Option Statistics			
Seq. No.	Scale -Item	Prop. Correct	Point Biser.	Alt.	Prop Endorsing	Point Biser.	Key
27	0-19	.46	.46	1	.17	-.20	
				2	.20	-.16	
				3	.46	.46	*
				4	.17	-.25	
				Other	.00		
28	0-20	.63	.44	1	.23	-.28	
				2	.12	-.22	
				3	.03	-.15	
				4	.63	.44	*
				Other	.00		
30	0-21	.55	.28	1	.26	-.12	
				2	.08	-.17	
				3	.11	-.13	
				4	.55	.28	*
				Other	.00		
33	0-22	.76	.27	1	.05	-.13	
				2	.76	.27	*
				3	.13	-.16	
				4	.06	-.14	
				Other	.00		
34	0-23	.65	.41	1	.16	-.25	
				2	.65	.41	*
				3	.08	-.31	
				4	.10	-.06	
				Other	.00	.03	
35	0-24	.76	.25	1	.07	-.14	
				2	.03	-.12	
				3	.14	-.14	
				4	.76	.25	*
				Other	.00		
37	0-25	.81	.31	1	.03	-.12	
				2	.05	-.20	
				3	.81	.31	*
				4	.10	-.19	
				Other	.00		
38	0-26	.79	.30	1	.07	-.16	
				2	.79	.30	*
				3	.05	-.14	
				4	.09	-.18	
				Other	.00		
39	0-27	.84	.35	1	.06	-.25	
				2	.05	-.13	
				3	.84	.35	*
				4	.05	-.19	
				Other	.00		

Appendix B Continued

		Item Statistics		Option Statistics			
Seq. No.	Scale -Item	Prop. Correct	Point Biser.	Alt.	Prop Endorsing	Point Biser.	Key
40	0-28	.60	.41	1	.60	.41	*
				2	.09	-.14	
				3	.22	-.30	
				4	.09	-.13	
41	0-29	.91	.32	1	.04	-.17	
				2	.03	-.15	
				3	.03	-.22	
				4	.91	.32	*
				Other	.00		
42	0-30	.58	.40	1	.17	-.12	
				2	.13	-.24	
				3	.12	-.23	
				4	.58	.40	*
				Other	.00		
44	0-31	.66	.43	1	.08	-.19	
				2	.66	.43	*
				3	.22	-.30	
				4	.04	-.15	
				Other	.00		
45	0-32	.72	.43	1	.11	-.24	
				2	.04	-.21	
				3	.72	.43	*
				4	.13	-.22	
				Other	.00		
48	0-33	.56	.33	1	.56	.33	*
				2	.16	-.16	
				3	.16	-.13	
				4	.13	-.17	
				Other	.00	-.02	
49	0-34	.76	.33	1	.07	-.14	
				2	.76	.33	*
				3	.11	-.13	
				4	.06	-.27	
				Other	.00		
50	0-35	.82	.41	1	.05	-.24	
				2	.07	-.23	
				3	.06	-.21	
				4	.82	.41	*
				Other	.00		
55	0-36	.51	.37	1	.51	.37	*
				2	.16	-.12	
				3	.30	-.25	
				4	.03	-.20	
				Other	.00		

Appendix B Continued

		Item Statistics		Option Statistics			
Seq. No.	Scale -Item	Prop. Correct	Point Biser.	Alt.	Prop Endorsing	Point Biser.	Key
56	0-37	.78	.32	1	.09	-.13	
				2	.09	-.21	
				3	.03	-.20	
				4	.78	.32	*
				Other	.00		
57	0-38	.74	.39	1	.74	.39	*
				2	.14	-.21	
				3	.06	-.24	
				4	.05	-.16	
				Other	.00		
58	0-39	.73	.18	1	.10	-.07	
				2	.73	.18	*
				3	.13	-.12	
				4	.04	-.10	
				Other	.00	-.00	
59	0-40	.56	.34	1	.56	.34	*
				2	.25	-.17	
				3	.07	-.12	
				4	.12	-.19	
				Other	.00	-.02	
61	0-41	.70	.38	1	.70	.38	*
				2	.06	-.15	
				3	.14	-.24	
				4	.09	-.19	
				Other	.00	-.01	
63	0-42	.75	.35	1	.03	-.10	
				2	.10	-.21	
				3	.75	.35	*
				4	.12	-.21	
				Other	.00	-.00	
64	0-43	.72	.45	1	.09	-.25	
				2	.06	-.24	
				3	.13	-.22	
				4	.72	.45	*
				Other	.00	-.02	
65	0-44	.84	.29	1	.84	.29	*
				2	.13	-.22	
				3	.02	-.14	
				4	.01	-.11	
				Other	.00		
66	0-45	.69	.38	1	.69	.38	*
				2	.05	-.17	
				3	.16	-.28	
				4	.11	-.13	
				Other	.00		

Appendix B Continued

		Item Statistics		Option Statistics			
Seq. No.	Scale -Item	Prop. Correct	Point Biser.	Alt.	Prop Endorsing	Point Biser.	Key
67	0-46	.77	.51	1	.05	-.26	
				2	.77	.51	*
				3	.09	-.27	
				4	.09	-.27	
				Other	.00		
68	0-47	.63	.28	1	.63	.28	*
				2	.04	-.19	
				3	.24	-.13	
				4	.09	-.15	
				Other	.00	.00	
69	0-48	.75	.33	1	.13	-.19	
				2	.07	-.14	
				3	.75	.33	*
				4	.05	-.20	
				Other	.00	.00	

Appendix C

Ratings of the 20 Best Answer English 30 Items: Five Judges

Appendix C: Ratings of the 20 Best Answer English 30 Items: Five Judges

item 5

A	B	C	D
1	3	4	2
1	2	3	4
1	2	3	4
1	2	4	3
1	2	4	3

item 8

4	3	2	1
1	3	4	2
1	3	4	2
1	2	3	4
1	3	4	2

item 12

4	3	2	1
2	3	1	4
3	4	1	2
3	4	1	2
2	4	1	3

item 14

1	2	4	3
1	2	3	4
1	3	2	4
1	2	3	4
1	2	3	4

item 16

3	1	4	2
4	1	2	3
3	1	4	2
3	1	4	2
4	1	2	3

item 18

4	1	2	3
4	3	1	2
4	3	1	2
3	2	1	4
3	4	1	2

item 21

4	3	1	2
2	3	4	1
2	3	4	1
3	4	2	1
2	3	4	

item 29

A	B	C	D
3	2	4	1
3	4	2	1
3	4	2	1
3	2	4	1
3	4	2	1

item 31

1	3	2	4
1	4	3	2
1	3	4	2
1	4	3	2
1	4	3	2

item 32

3	4	1	2
3	4	1	2
3	4	1	2
4	3	1	2
3	4	1	2

item 36

2	1	4	3
2	1	3	4
2	1	3	4
2	1	3	4
2	1	3	4

item 46

2	4	3	1
4	2	3	1
2	4	3	1
3	2	4	1
4	2	3	1

item 47

2	3	4	1
2	4	3	1
2	4	3	1
2	3	4	1
2	4	3	1

item 51

1	4	3	2
3	2	1	4
2	4	1	3
2	4	1	3
3	2	1	4

item 52

A	B	C	D
4	3	1	2
2	4	1	3
3	2	1	4
4	2	1	3
2	4	1	3

item 53

4	2	3	1
4	3	1	2
3	4	1	2
4	3	1	2
4	3	1	2

item 54

4	2	1	3
2	1	3	4
2	1	4	3
	1		
2	1	4	3

item 60

3	2	1	4
4	2	1	3
2	3	1	4
		1	2
4	2	1	3

item 62

2	1	3	4
3	1	2	4
2	1	4	3
3	1	4	2
3	1	2	4

item 70

1	4	3	2
1	4	2	3
1	3	2	4
1	4	3	2
1	4	2	3

Appendix D

Conventional Item Analysis: ENGBD Subtest

Appendix D: Conventional Item Analysis: ENGBD Subtest

		Item Statistics		Option Statistics			
Seq. No.	Scale -Item	Prop. Correct	Point Biser.	Alt.	Prop Endorsing	Point Biser.	Key
5	1-1	.67	.40	1	.67	.40	*
				2	.23	-.27	
				3	.04	-.15	
				4	.05	-.19	
				Other	.00	.03	
8	1-2	.57	.45	1	.57	.45	*
				2	.13	-.15	
				3	.04	-.14	
				4	.25	-.32	
				Other	.00		
12	1-3	.51	.40	1	.19	-.30	
				2	.09	-.20	
				3	.51	.40	*
				4	.21	-.07	
				Other	.00		
14	1-4	.66	.35	1	.66	.35	*
				2	.22	-.20	
				3	.06	-.19	
				4	.06	-.15	
				Other	.00	-.04	
16	1-5	.75	.37	1	.05	-.12	
				2	.75	.37	*
				3	.08	-.16	
				4	.12	-.29	
				Other	.00	.00	
18	1-6	.51	.36	1	.14	-.10	
				2	.14	-.23	
				3	.51	.36	*
				4	.21	-.16	
				Other	.00		
21	1-7	.70	.38	1	.12	-.12	
				2	.10	-.25	
				3	.08	-.22	
				4	.70	.38	*
				Other	.00	-.02	
29	1-8	.69	.32	1	.07	-.12	
				2	.10	-.20	
				3	.14	-.17	
				4	.69	.32	*
				Other	.00		
31	1-9	.80	.31	1	.80	.31	*
				2	.01	-.14	
				3	.03	-.17	
				4	.15	-.22	
				Other	.00		

Appendix D Continued

		Item Statistics		Option Statistics			
Seq. No.	Scale -Item	Prop. Correct	Point Biser.	Alt.	Prop Endorsing	Point Biser.	Key
32	1-10	.71	.35	1	.05	-.12	
				2	.08	-.25	
				3	.71	.35	*
				4	.15	-.18	
				Other	.00		
36	1-11	.57	.42	1	.25	-.20	
				2	.57	.42	*
				3	.13	-.28	
				4	.05	-.12	
				Other	.00	-.04	
46	1-12	.73	.44	1	.09	-.23	
				2	.08	-.26	
				3	.10	-.20	
				4	.73	.44	*
				Other	.00		
47	1-13	.76	.40	1	.11	-.26	
				2	.04	-.20	
				3	.09	-.18	
				4	.76	.40	*
				Other	.00		
51	1-14	.46	.37	1	.16	-.13	
				2	.20	-.21	
				3	.46	.37	*
				4	.18	-.14	
				Other	.00	-.02	
52	1-15	.89	.30	1	.04	-.20	
				2	.04	-.15	
				3	.89	.30	*
				4	.04	-.15	
				Other	.00	-.02	
53	1-16	.44	.45	1	.08	-.25	
				2	.15	-.14	
				3	.44	.45	*
				4	.33	-.22	
				Other	.00	-.01	
54	1-17	.57	.33	1	.12	-.20	
				2	.57	.33	*
				3	.16	-.12	
				4	.15	-.16	
				Other	.00	.00	
60	1-18	.59	.33	1	.09	-.18	
				2	.25	-.15	
				3	.59	.33	*
				4	.07	-.17	
				Other	.00		

Appendix D Continued

		Item Statistics		Option Statistics			
Seq. No.	Scale -Item	Prop. Correct	Point Biser.	Alt.	Prop Endorsing	Point Biser.	Key
62	1-19	.72	.28	1	.11	-.10	
				2	.72	.28	*
				3	.16	-.23	
				4	.02	-.07	
				Other	.00	-.02	
70	1-20	.68	.36	1	.68	.36	*
				2	.04	-.19	
				3	.13	-.18	
				4	.14	-.20	
				Other	.00	.01	

Appendix E

Conventional Item Analysis: TTWI Subtest

Appendix E: Conventional Item Analyses: TTWI Subtest

		Item Statistics		Option Statistics			
Seq. No.	Scale -Item	Prop. Correct	Point Biser.	Alt.	Prop Endorsing	Point Biser.	Key
2	1-1	.92	.24	1	.01	-.09	
				2	.92	.24	*
				3	.03	-.13	
				4	.04	-.17	
				Other	.00		
3	1-2	.70	.15	1	.12	-.14	
				2	.03	-.09	
				3	.70	.15	*
				4	.14	-.02	
				Other	.00		
5	1-3	.53	.24	1	.18	-.16	
				2	.53	.24	*
				3	.16	-.06	
				4	.13	-.11	
				Other	.00	-.02	
7	1-4	.68	.22	1	.24	-.13	
				2	.68	.22	*
				3	.02	-.08	
				4	.06	-.15	
				Other	.00	-.03	
8	1-5	.56	.22	1	.56	.22	*
				2	.31	-.10	
				3	.03	-.10	
				4	.09	-.15	
				Other	.00	-.02	
9	1-6	.44	.24	1	.05	-.16	
				2	.44	.24	*
				3	.33	-.08	
				4	.17	-.11	
				Other	.00	-.05	
11	1-7	.88	.25	1	.88	.25	*
				2	.03	-.15	
				3	.03	-.14	
				4	.07	-.13	
				Other	.00		
13	1-8	.90	.23	1	.02	-.11	
				2	.07	-.16	
				3	.90	.23	*
				4	.01	-.13	
				Other	.00		
14	1-9	.39	.25	1	.31	-.12	
				2	.07	-.12	
				3	.39	.25	*
				4	.22	-.08	
				Other	.00	-.01	

Appendix E Continued

		Item Statistics		Option Statistics			
Seq. No.	Scale -Item	Prop. Correct	Point Biser.	Alt.	Prop Endorsing	Point Biser.	Key
17	1-10	.48	.26	1	.03	-.19	
				2	.35	-.12	
				3	.48	.26	*
				4	.14	-.11	
				Other	.00	-.00	
19	1-11	.55	.22	1	.55	.22	*
				2	.17	-.10	
				3	.14	-.13	
				4	.14	-.06	
				Other	.00	-.01	
22	1-12	.49	.28	1	.24	-.07	
				2	.49	.28	*
				3	.21	-.23	
				4	.06	-.05	
				Other	.00		
24	1-13	.68	.29	1	.68	.29	*
				2	.09	-.17	
				3	.15	-.09	
				4	.08	-.20	
				Other	.00		
25	1-14	.72	.22	1	.05	-.09	
				2	.12	-.17	
				3	.72	.22	*
				4	.11	-.08	
				Other	.00	.00	
28	1-15	.33	.34	1	.21	-.10	
				2	.19	-.10	
				3	.27	-.18	
				4	.33	.34	*
				Other	.00		
29	1-16	.69	.37	1	.11	-.18	
				2	.16	-.23	
				3	.04	-.16	
				4	.69	.37	*
				Other	.00	.02	
30	1-17	.26	.19	1	.30	-.10	
				2	.26	.19	*
				3	.28	.01	
				4	.16	-.10	
				Other	.00		
31	1-18	.54	.24	1	.54	.24	*
				2	.25	-.08	
				3	.16	-.15	
				4	.06	-.12	
				Other	.00	-.04	

Appendix E Continued

		Item Statistics		Option Statistics			
Seq. No.	Scale -Item	Prop. Correct	Point Biser.	Alt.	Prop Endorsing	Point Biser.	Key
33	1-19	.57	.39	1	.04	-.08	
				2	.57	.39	*
				3	.06	-.12	
				4	.33	-.31	
				Other	.00		
35	1-20	.59	.32	1	.24	-.20	
				2	.59	.32	*
				3	.11	-.14	
				4	.05	-.12	
				Other	.00		
36	1-21	.31	.23	1	.27	.05	
				2	.31	.23	*
				3	.26	-.25	
				4	.16	-.06	
				Other	.00		
39	1-22	.19	.25	1	.19	.25	*
				2	.17	-.08	
				3	.23	-.12	
				4	.41	-.04	
				Other	.00	-.02	
40	1-23	.65	.34	1	.08	-.18	
				2	.65	.34	*
				3	.08	-.17	
				4	.19	-.16	
				Other	.00	-.08	
41	1-24	.55	.21	1	.11	-.07	
				2	.25	-.17	
				3	.55	.21	*
				4	.09	-.03	
				Other	.00	-.01	
42	1-25	.44	.44	1	.44	.44	*
				2	.08	-.17	
				3	.33	-.23	
				4	.14	-.18	
				Other	.00	.03	
44	1-26	.56	.35	1	.18	-.15	
				2	.12	-.17	
				3	.13	-.17	
				4	.56	.35	*
				Other	.00	.00	
45	1-27	.34	.21	1	.13	-.12	
				2	.24	-.12	
				3	.28	-.02	
				4	.34	.21	*
				Other	.00		

Appendix E Continued

		Item Statistics		Option Statistics			
Seq. No.	Scale -Item	Prop. Correct	Point Biser.	Alt.	Prop Endorsing	Point Biser.	Key
46	1-28	.37	.22	1	.14	.06	
				2	.31	-.19	
				3	.18	-.11	
				4	.37	.22	*
				Other	.00	.01	
47	1-29	.63	.33	1	.11	-.15	
				2	.14	-.16	
				3	.12	-.18	
				4	.63	.33	*
				Other	.00		
48	1-30	.66	.32	1	.08	-.18	
				2	.14	-.17	
				3	.12	-.12	
				4	.66	.32	*
				Other	.00		
49	1-31	.33	.24	1	.20	-.02	
				2	.33	.24	*
				3	.39	-.16	
				4	.06	-.11	
				Other	.01	-.03	
50	1-32	.75	.34	1	.03	-.13	
				2	.09	-.23	
				3	.75	.34	*
				4	.13	-.17	
				Other	.01	-.03	

Appendix F

Conventional and Item Response Item Analyses for TTWI Subtest: Removal of 71 Examinees with Chance Scores

Appendix F: Conventional and Item Response Item Analyses for TTWI Subtest: Removal of 71 Examinees with Chance Scores

	Conventional			IRT Parameter Estimates					
				1PL	2PL		3PL		
Item	p-value	pbs	option	b	b	a	b	a	c
2	0.93*	0.21	a,c,d	-6.12*	-3.68*	0.44	-3.09*	0.46	0.25
3	0.71	0.13*	b	-2.16	-3.58	0.15	-0.58	0.18	0.36
5	0.54	0.22	✓	-0.38	-0.46	0.20	2.02*	0.53	0.44
7	0.69	0.19*	c	-1.91	-2.29	0.21	-0.31	0.26	0.32
8	0.57	0.19*	c	-0.74	-0.96	0.19	1.72	0.27	0.37
9	0.44	0.24	✓	0.54	0.54	0.25	1.91	0.77	0.36
11	0.89	0.22	b,c	-4.94*	-2.91	0.45	-2.14*	0.53	0.26
13	0.91*	0.19*	a,d	-5.68*	-3.30	0.45	-2.55	0.52	0.25
14	0.40	0.24	✓	1.08	1.09	0.25	2.16	0.68	0.31
17	0.49	0.25	a	0.13	0.12	0.26	1.58	0.96	0.40
19	0.56	0.20	✓	-0.58	-0.72	0.20	1.89	0.88	0.50
22	0.50	0.27	✓	0.08	0.07	0.27	1.65	0.84	0.40
24	0.69	0.27	✓	-1.91	-1.59	0.31	0.12	0.42	0.34
25	0.72	0.19*	✓	-2.33	3.01	0.19	-1.16	0.21	0.31
28	0.34	0.34	✓	1.64*	0.97	0.45	1.36	0.78	0.17
29	0.70	0.36	c	-2.07	-1.11	0.52	-0.52*	0.62	0.23
30	0.26	0.19*	✓	2.63	2.60*	0.25	2.02	1.34	0.21
31	0.55	0.21	✓	-0.51*	-0.60	0.21	2.02	0.49	0.45
33	0.58	0.38	a	-0.72*	-0.38*	0.54	0.37*	0.86	0.28
35	0.61	0.30	✓	-1.04	-0.76	0.36	0.69	0.69	0.39
36	0.31	0.23	✓	1.88	1.74	0.27	2.26	0.63	0.23
39	0.20	0.26	✓	3.52	2.41	0.38	1.93	1.25	0.14
40	0.66	0.33	✓	-1.59*	-0.97	0.44	-0.08	0.59	0.29

Item	Conventional			IRT Parameter Estimates					
				1PL	2PL		3PL		
	p-value	pbs	option	b	b	a	b	a	c
41	0.55	0.19*	✓	-0.50	-0.69	0.18	2.42	0.43	0.46
42	0.46	0.43	✓	0.50*	0.22*	0.63	0.62*	0.80	0.16
44	0.58	0.35	✓	-0.71*	-0.43*	0.46	0.50*	0.79	0.31
45	0.35	0.22	✓	1.54	1.64	0.23	2.32	0.69	0.28
46	0.38	0.21	✓	1.30	1.54	0.21	2.56	0.69	0.30
47	0.64	0.32	✓	-1.40*	-0.92	0.40	0.08	0.55	0.29
48	0.67	0.29	✓	-1.79*	-1.22	0.39	-0.01	0.56	0.34
49	0.34	0.24	✓	1.72	1.75	0.25	2.32	0.75	0.27
50	0.76	0.32	a	-2.84*	-1.53*	0.51	-0.77*	0.62	0.29

Appendix G

Conventional Item Analysis: TTWDOS Subtest

Appendix G: Conventional Item analysis: TTWDOS Subtest

		Item Statistics		Option Statistics			
Seq. No.	Scale -Item	Prop. Correct	Point Biser.	Alt.	Prop Endorsing	Point Biser.	Key
1	0-1	.67	.24	1	.67	.24	*
				2	.12	-.17	
				3	.10	-.12	
				4	.11	-.06	
				Other	.00	-.06	
4	0-2	.33	.33	1	.17	-.18	
				2	.33	.33	*
				3	.04	-.06	
				4	.46	-.14	
				Other	.00		
6	0-3	.40	.28	1	.40	.28	*
				2	.14	-.15	
				3	.35	-.11	
				4	.10	-.10	
				Other	.00	-.02	
10	0-4	.35	.25	1	.35	.25	*
				2	.34	-.12	
				3	.17	-.11	
				4	.13	-.06	
				Other	.00	.01	
12	0-5	.58	.24	1	.06	-.12	
				2	.24	-.07	
				3	.11	-.19	
				4	.58	.24	*
				Other	.00	-.01	
15	0-6	.50	.31	1	.11	-.17	
				2	.50	.31	*
				3	.22	-.09	
				4	.16	-.18	
				Other	.00	.01	
16	0-7	.39	.28	1	.09	-.11	
				2	.26	-.13	
				3	.39	.28	*
				4	.25	-.11	
				Other	.00	-.02	
18	0-8	.67	.34	1	.08	-.19	
				2	.12	-.07	
				3	.13	-.25	
				4	.67	.34	*
				Other	.00	.03	
20	0-9	.27	.25	1	.27	.25	*
				2	.16	-.12	
				3	.27	-.09	
				4	.30	-.05	
				Other	.00	-.03	

Appendix G Continued

		Item Statistics		Option Statistics			
Seq. No.	Scale -Item	Prop. Correct	Point Biser.	Alt.	Prop Endorsing	Point Biser.	Key
21	0-10	.58	.34	1	.23	-.13	
				2	.09	-.20	
				3	.10	-.19	
				4	.58	.34	*
				Other	.00	.00	
23	0-11	.29	.21	1	.19	-.04	
				2	.29	.21	*
				3	.26	-.07	
				4	.26	-.12	
				Other	.00	-.03	
26	0-12	.63	.23	1	.11	-.04	
				2	.63	.23	*
				3	.17	-.20	
				4	.08	-.06	
				Other	.00	-.06	
27	0-13	.39	.35	1	.29	-.15	
				2	.39	.35	*
				3	.15	-.10	
				4	.17	-.17	
				Other	.00	-.04	
32	0-14	.42	.23	1	.42	.23	*
				2	.21	-.09	
				3	.23	-.12	
				4	.13	-.08	
				Other	.00	-.01	
34	0-15	.61	.29	1	.07	-.15	
				2	.61	.29	*
				3	.16	-.10	
				4	.17	-.19	
				Other	.00	.03	
37	0-16	.31	.27	1	.21	-.12	
				2	.15	-.15	
				3	.33	-.04	
				4	.31	.27	*
				Other	.00	.00	
38	0-17	.50	.35	1	.50	.35	*
				2	.07	-.20	
				3	.33	-.19	
				4	.11	-.12	
				Other	.00		
43	0-18	.59	.26	1	.24	-.13	
				2	.06	-.06	
				3	.59	.26	*
				4	.12	-.18	
				Other	.00		

Appendix H

Conventional and Item Response Item Analyses for TTWDOS Subtest: Removal of 71 Examinees with Chance scores

Appendix H: Conventional and Item Response Item Analyses for TTWDOS Subtest: Removal of 71 Examinees with Chance scores

	Conventional			IRT Parameter Estimates					
				1PL	2PL		3PL		
Item	p-value	r-pbis	option	b	b	a	b	a	c
1	0.68	0.20	✓	-3.78	-3.01	0.16	0.01	0.20	.38
4	0.34	0.31	c	3.12	1.16	0.35	1.60	0.85	.23
6	0.41	0.28	✓	1.78	0.88	0.26	1.95	1.18	.36
10	0.37	0.21	✓	2.59	1.88	0.17	3.31	0.82	0.35
12	0.60	0.20	✓	-1.98	-1.72	0.14	3.06	0.29	.50
15	0.52	0.28	✓	-0.38	-0.22	0.22	2.38	0.77	.47
16	0.41	0.24	✓	1.75	1.24	0.18	3.54	0.97	.40
18	0.69	0.30	✓	-3.90	-1.40	0.37	-0.40	0.51	.27
20	0.28*	0.24	✓	4.56	2.45	0.23	2.77	0.90	.25
21	0.60	0.31	✓	-1.99	-0.80	0.32	0.36	0.54	.30
23	0.30	0.21	✓	4.10	2.65	0.19	3.72	0.88	.29
26	0.65	0.19*	✓	-3.04	-2.28	0.16	1.00	0.34	.44
27	0.40	0.33	✓	1.88	0.69	0.35	1.55	0.88	.29
32	0.44	0.21	✓	1.21	0.93	0.16	2.82	1.15	.42
34	0.63	0.25	✓	-2.57	-1.30	0.25	0.35	0.42	.33
37	0.31	0.28	✓	3.79	1.78	0.27	2.13	0.99	.26
38	0.52	0.33	✓	-0.31	-0.13	0.33	0.99	0.60	0.31
43	0.61	0.23	✓	-2.12	-1.47	0.18	2.38	0.36	0.50